Supplementary Note

Analysis of Illumina methylation arrays

Custom GoldenGate Array

We selected 151 of the cDMRs identified in Irizarry et al.[1], regions consistently differentially methylated in all 13 colon cancers studied by comprehensive high-throughput array based methylation (CHARM) analysis. Probes were designed around CpGs that showed consistent differences in CHARM, while passing Illumina's quality control metrics[2]. The resulting probes covered 139 regions, with 1-7 probes per region. The majority of the probes were in CpG island shores (66%), defined as less than 2 kb away from the edge of a canonically defined high-CpG density island[1]. The remainder of the probes were either inside CpG islands (11%) or >2 kb distant (23%).

Sample Preparation

Cryogenically stored freshly frozen samples were obtained from the Cooperative Human Tissue Network (NCI, Bethesda, MD), the National Wilms Tumor Study tissue bank (Edmonton, Alberta, Canada) and the Johns Hopkins Hospital, under an IRB-approved waiver of consent. In total 290 samples were assayed, including cancers from colon (10), lung (24), breast (27), thyroid (36), and kidney (Wilms') (25), with matched normal tissues to 111 of these 122 cancers, along with 30 colon premalignant adenomas, 18 normal colon, and 9 normal breast samples. Two small sections were taken from each sample (~0.5cmx0.5cmx0.2cm); one for DNA purification and one for histopathology. Histopathology samples were submitted to the immunohistochemistry lab at Johns Hopkins Hospital for processing. Normal and cancer samples were matched from the same patient and the same tissue whenever available.

A board-certified oncology pathologist validated classification of all samples independently and blindly. The pathologist also quantified specific cellular subtypes and p53 status for tumor and normal specimens from colon and kidney. Supplementary Figure 3 summarizes the histological analysis of the colon and kidney normal and tumor samples demonstrating that normal samples are typically more heterogeneous in cellular composition than the tumor samples.

DNA purification was done using either the DNeasy Blood and Tissue Kit (Qiagen) or the MasterPure Complete DNA Purification Kit (Epicenter). DNA concentration and purity was assayed using a Nanodrop spectrophotometer. Tumor and normal status and tissue type were balanced by plate to avoid batch effects. Methylated and unmethylated controls (Zymo Research), along with sample cross-plate controls, were included on plates. Samples were bisulfite treated using the EZ-96 Methylation Gold kit (Zymo Research), and hybridization was performed by the Center for Inherited Disease Research of Johns Hopkins University.

Custom Illumina methylation array processing and analysis

We quantile normalized[3] separately the raw intensity data from the Cy5 and Cy3 channels representing methylated and unmethylated DNA, and methylation level was calculated as the ratio of the Cy5 intensity over the sum of the intensities from both channels. To control for array quality, arrays for which the average of the median log intensities from the two channels was small (<7), or for which the median absolute deviation of the overall methylation signal was small (<1.9) were removed from the dataset. We ruled out batch effects following the procedures described by Leek et al. [4] Differences in methylation variability were measured and tested using an F-test. Differences in mean methylation levels were measured and tested using a t-test. Significance was taken as 0.01.

Age Regression of Custom Illumina methylation array

Because not all our samples were paired and age is known to affect methylation profiles, we ran an analysis to verify that our results were not being driven by patient age. For this, we only used the tissues for which we had sufficient age data available (lung, breast and thyroid). We did not include Wilms' tumor in our analysis since it is a childhood disease. We then regressed age out of our methylation data by fitting a linear model of the form: $logit(\beta_j) = \mu + \alpha \cdot age_j + \varepsilon$ for each CpG, where β_j is the measured methylation level for sample *j*, μ is an over-all mean methylation estimate, age_j is the age of sample *j* and ε is an error term. Using the age-corrected measurement $inverse\text{-}logit\ (\ logit(\beta_j) - \alpha \cdot age_j$), we repeated the analysis shown in Figure 1, plotting a per CpG analysis of methylation values in cancer versus normal (Supplementary Fig. 4a-c) and a cluster analysis of the normal tissues using the 25 most variable probes (Supplementary Fig 4d). We obtained almost identical results.

We also wanted to determine if methylation level had a causal role in the increased variation observed in cancer, i.e. greater variability at 50% methylation due to the binomial nature of methylation. To test this, we used an overdispersed binomial model[5]. We then plotted for each CpG the mean versus standard deviation of methylation values for normal and cancer in colon, lung, breast, thyroid and kidney (Wilms' tumor). The dotted line indicates the expected variance from the binomial model at each mean methylation level. Increased variability is clearly observed in cancer along the range of methylation values. CpGs are coded to indicate significant differences in mean only (green), variance only (purple) or both (orange). From this analysis we concluded that methylation level was not the main factor in the observed increase variation.

Selected sample copy number analysis

We obtained the raw data from existing CHARM experiments[1], which included five of the colon cancer samples hybridized on the custom GoldenGate array. We then hybridized five of the Wilms tumor samples, along with normal kidney samples as controls, to the CHARM array. The 5 colon cancer samples had matched normal sample unlike the 5 WIlms tumor samples where we used 5 normal Kidney tissue as controls. We extracted the intensities from the Cy3 channel, which corresponds to total DNA. These intensities were corrected for spatial effect, quantile normalized, and corrected for sequence effect[6]. Log-ratios were formed comparing the intensities from the cancer to intensities for normal samples and the median value of each CHARM region[7]. Each of the CHARM regions was then computed. We then used the circular binary segmentation (CBS) algorithm[8] to find regions of copy number variation. The resulting log ratios and estimated copy number segments are shown in Supplementary Figure 5, which shows a high degree of copy number variation in colon compared to Wilms.

Illumina HumanMethylation27k array analysis

We downloaded a publicly available dataset of methylation levels of 22 matched colon normal/tumor samples assayed using Illumina's HumanMethylation27K array (Gene Expression Omnibus accession number GSE17648). Probes were annotated according to their genomic distance from the edge of canonically defined CpG islands[1]: 42% inside islands, 31.6% in CpG island shores, and the remaining 26.4% were >2 kb distant. The methylation measurements were used with no further preprocessing. Differences in methylation variability were measured and tested using an F-test. Differences in mean methylation levels were measured and tested using a t-test. Significance was taken as 0.01.

Analysis of whole genome and capture bisulfite sequencing

Whole genome bisulfite sequencing

Bisulfite sequencing libraries were prepared using the approach previously described by Bormann Chung *et al.*[9], with some modifications so the entire protocol is provided in detail here. 5ug of genomic DNA was sheared using a Covaris E2 sonicator. 40ng of AluI-digested unmethylated λ DNA was added to each

sample in order to monitor the efficiency of bisulfite conversion. Sample DNA ends were then repaired using 1x End Polish Buffer, 400nM each of dATP, dGTP and dTTP (leaving out dCTP), 40 U of End Polishing Enzyme 1 (Life Tech) and 80 U of End Polishing Enzyme 2 (Life Tech). Adaptor sequences, as detailed by Bormann Chung *et al.* [9], were then ligated onto the samples, using 1X T4 ligase buffer, 4.5uM of methyl-protect P1 adaptor, 4.5uM of P2 adaptor and 50 U of T4 ligase. Nick translation was performed in 1X Exo-Klenow buffer, 0.5mM dNTPs containing methyl dCTP and 20 units of Exo-Klenow-Fragment (Ambion) for 1 hour at 16 ^{o}C. 500ng aliquots of the resulting product was then bisulfite converted using formamide as an additional denaturant[10]. 24µL of formamide was added to an equal volume of DNA and incubated at 95^{o}C for 5 minutes. Subsequently, 100µL of Zymo Gold bisulfite conversion reagent (Zymo) was added, and the mixture was incubated for 8 hours in 50^{o}C. Samples were then desulphonated and purified using spin columns following the EZ-DNA Zymo Methylation-Gold protocol. 5µL of the bisulfite converted library was amplified in 1X PCR buffer, 0.2 mM dNTP, 1mM each of the standard SOLiD fragment library primers, 5 U of Taq (Denville), and 0.25 U of Pfu Turbo Taq (Stratagene). The DNA was subjected to 8 cycles of PCR and the resulting product was purified using AMpure SPRI beads (Beckman Genomics). The libraries were then sequenced on the SOLiD 3+ platform yielding 50 base pair reads.

Capture bisulfite sequencing

A BSPP library was custom designed based in part on DMRs previously found in colon cancer[1], covering ~60,000 highly curated differentially methylated regions in the human genome (620,708 CpG sites in 19.2Mb of genomic regions covered). 1µg of genomic DNA from the three tumor-normal samples was bisulfite converted using the Zymo EZ Methylation Gold Kit. The bisulfite converted DNA was then captured with BSPP using a previously described method[11]. Briefly, 300ng of bisulfite converted DNA was captured in a 10µl reaction containing 1X Ampligase Buffer (Epicenter Biotechnologies). The mixture was heated to 95˚C for 30 seconds and then incubated for 20 hours at 58˚C. A 2µl mix containing 2U/µl AmpliTaq Stoffel fragment, 0.5U/µl Ampligase, 50µM dNTP was added to each tube. The tubes were incubated at 58˚C for another 28 hours and then heat inactivated at 94˚C for 2 minutes. Single stranded DNA was removed by adding 2µl of Exo I/III mix and incubating at 37˚C for 1 hour. 300ng of the captured padlock probes were amplified as previously described and then digested with Mme I in order to remove the amplification primers[11]. Adapters compatible with the Illumina GAII sequencer were then ligated to the digested DNA in order to generate the sequencing libraries[11].

Alignment of sequencing reads from bisulfite treated DNA

We developed a custom alignment tool for Illumina and SOLiD sequencing reads derived from bisulfite-treated DNA. The tool aligned reads with the aid of a spaced-seed index of the genome while biasing neither toward nor against methylated cytosines in CpGs. Note that aligners can introduce a bias when an unmethylated C (which becomes a T) is penalized for aligning to a C in the genome, but a methylated C (which remains a C) is not. The opposite bias can also occur, e.g., if all CpGs are converted to TpGs in the reference prior to alignment. Other projects address this in part by additionally converting Cs to Ts in the reads[12]. But this approach is not applicable to the colorspace reads generated by the SOLiD instrument, for which nucleotide positions encoding Cs cannot be accurately determined prior to alignment. The aligner used here leaves each read as-is but penalizes neither C-to-C nor T-to-C partial alignments in CpGs.

For alignment we extend the approach taken by the BSMAP tool[13]. Our approach supports a broad range of spaced-seed designs and extends the BSMAP approach to allow alignment of SOLiD colorspace reads as well as typical Illumina reads. Like in BSMAP, C/T bias is avoided by creating and storing multiple copies (potentially) of each reference subsequence indexed, one copy for each distinct assignment of Cs and Ts to genomic Cs or CpGs present in the subsequence. For colorspace reads, our

algorithm extracts subsequences of colors (rather than nucleotides), and a copy is created for every distinct assignment of Cs and Ts to genomic Cs present in any nucleotide overlapped by any color in the extracted subsequence.

For these experiments, the alignment algorithm is configured to remove the penalty associated with either a C or a T aligning to a C in a CpG, and to treat non-CpG Cs in the genome as Ts. This policy removes bias from CpG methylation measurements, but assumes that there is little or no non-CpG cytosine methylation. Where non-CpG cytosine methylation occurs, this approach is more likely to fail to find alignments overlapping the methylated cytosine, and the resulting consensus will contain an anti-methylation bias at the methylated cytosine.

The algorithm was extended to handle data both from protocols that yield sequencing reads only from the bisulfite-treated Watson and Crick strands (as is the case for the whole-genome bisulfite SOLiD sequencing data discussed here), as well as from protocols that yield those sequences and their reverse complements (as is the case for the capture bisulfite Illumina sequencing data).

Alignment of SOLiD sequencing reads from whole-genome bisulfite-treated DNA: The algorithm described above was used to align a total of 7.79 billion reads obtained from 8 runs of a SOLiD 3 Plus instrument against a reference sequence collection consisting of the GRCh37 human genome assembly (including mitochondrial DNA and "unplaced" contigs) plus the sequence of the spiked-in λ phage genome. Alignment was performed with respect to the bisulfite-treated Watson and Crick strands but not their reverse complements, per the sequencing protocol used. Each read obtained from the SOLiD 3 Plus instrument consists of a primer nucleotide followed by a string of 50 "colors," where each color encodes a class of dinucleotides according to the SOLiD colorspace encoding scheme. Prior to alignment, the initial primer base and 5'-most color were trimmed from all input reads, yielding a string of 49 colors. The alignment policy was selected to guarantee that all alignments with up to 3 color mismatches would be found, and some but not all alignments with 4-6 color mismatches would be found. The alignment of a T or C to a C in a CpG does not incur a mismatch penalty (except in some cases where a sequencing error is also present). The policy was also set to distinguish between reads that align uniquely and those that align non-uniquely. Color-to-color alignments are decoded into nucleotide alignments with a Viterbi-like algorithm[14]. The final alignment, when expressed in nucleotides, is one character shorter than the input read, i.e. 48 nucleotides long. Alignments for reads aligning non-uniquely are ignored in subsequent stages. Alignment results are summarized in Supplementary Table 14 and Supplementary Figure 19.

Alignment of Illumina sequencing reads from captured bisulfite-treated DNA: The algorithm described above was also used to align a total of 79.3 million reads obtained from an Illumina GA II instrument against a reference sequence collection consisting of the GRCh37 human genome assembly (including mitochondrial DNA and "unplaced" contigs). Alignment was performed with respect to the bisulfite-treated Watson and Crick strands and their reverse complements, per the sequencing protocol used. Each read consisted of either 73 nucleotides (for 14.5 million reads) or 80 nucleotides (for 64.8 million reads). No trimming was performed prior to alignment. The alignment policy was selected to guarantee that all alignments with up to 4 nucleotide mismatches would be found, and some but not all alignments with 5 or 6 nucleotide mismatches would be found. The alignment of a T or C to a C in a CpG does not incur a mismatch penalty (except in some cases where a sequencing error is also present). The policy was also set to distinguish between reads that align uniquely and those that align non-uniquely. Alignments for reads aligning non-uniquely are ignored in subsequent stages. Alignment results are summarized in Supplementary Table 15 and Supplementary Figure 21.

Extraction of methylation evidence from alignments: After alignment, a series of scripts extracted and summarized CpG methylation evidence present in the unique alignments. The evidence was compiled

into a set of per-sample, per-chromosome evidence tables. Alignments to the λ phage genome were also compiled into a separate table. A piece of CpG "evidence" was created when an alignment overlapped the cytosine position of a CpG in the reference and the overlapping nucleotide in the alignment was either a T (indicating a lack of methylation) or a C (indicating presence of methylation). Once a piece of evidence was extracted from a unique alignment, it was subjected to a filter. In the case of the SOLiD reads obtained by sequencing whole-genome bisulfite-treated DNA, the filter removed evidence that was either refuted by one or both of the overlapping colors from the original read, or was within 4 positions of either end of the nucleotide alignment. In the case of the Illumina reads obtained by sequencing captured, bisulfite-treated DNA, evidence within 15 positions of the beginning (5' end) of the read was discarded. The positions filtered in this step were determined by examining the M-bias lines (see below).

All evidence that passed the filtering step was added to the CpG summary table. A record in the table summarizes, for a given CpG: the filtered evidence nucleotides that aligned to it, the filtered quality values (i.e. of the two colors overlapping the evidence nucleotide for SOLiD data, or the of the overlapping nucleotide for Illumina data) that aligned to it, the number of distinct alignment positions from which filtered evidence was taken, the "mapability" of the CpG and surrounding bases (i.e. the number of 50-mers overlapping the CpG that are unique up to 3 mismatches), (e) the local CG content of the bases surrounded the CpG. The mapability measure for each genome position was pre-calculated using Bowtie[15]. Supplementary Tables 16 and 17 summarize the amount and type of evidence extracted at each stage for the whole-genome SOLiD bisulfite sequencing and Illumina capture bisulfite sequencing data respectively. Supplementary Table 18 summarizes the whole-genome SOLiD bisulfite sequencing CpG evidence coverage with respect to the GRCh37 human genome assembly for each sample. Finally, Supplementary Table 1 summarizes per-sample average coverage both genome-wide and for CpG cytosines for the whole genome bisulfite data and Supplementary Table 2 for the capture bisulfite sequencing data.

In the case of the SOLiD reads obtained by sequencing whole-genome bisulfite-treated DNA, evidence from reads that aligned uniquely to the λ genome was used to estimate the bisulfite conversion rate for unmethylated cytosines. The conversion rate was estimated as the fraction of high-quality evidence from reads aligning uniquely to the λ phage genome that indicated lack of methylation. Supplementary Figure 20 and the final column of Supplementary Table 14 show the estimates, which all lie between 99.7% and 99.8%.

To measure global prevalence of non-CpG cytosine methylation, we examined all filtered nucleotide evidence from the SOLiD reads overlapping non-CpG cytosine positions in the human reference genome. Filtered nucleotide evidence consists of evidence (a) from reads that aligned uniquely, (b) where both overlapping colors from the original read agree with the decoded nucleotide and, (c) where nucleotides within 4 positions of either end of the alignment are excluded. For each subject we measure the overall fraction of evidence at CpG cytosine positions where the overlapping nucleotide is a T or a C. We do the same for non-CpG cytosine positions. Supplementary Table 19 summarizes the results, comparing them with the rate of cytosine non-conversion estimated from filtered evidence aligning to the λ phage genome. We observe that for all subjects, the fraction of Cs observed overlapping non-CpG cytosines does not rise above the approximate fraction expected from unconverted cytosines.

Supplementary Figures 22 and 23 shows the results of a diagnostic assessing a type of bias in the filtered evidence from the reads that aligned uniquely to the human genome. The diagnostic is to calculate, for each offset from the 5' end of the read, the proportion of filtered evidence taken from any read at that position that indicates that methylation is present. This is the "M-bias line." One might expect this proportion to be independent of position, and therefore might expect the M-bias line to be flat and

horizontal. In practice, sequencing error and other noise arising from sample preparation and alignment cause the M-bias line to bend. In our experience, bends usually occur toward one or both ends of the read. The positional filtering criteria described above were designed to eliminate evidence from positions where the M-bias line deviated substantially from the main horizontal line. The relative flatness of the lines obtained for our samples after filtering gives us some additional assurance that the signal we obtain is not substantially affected by noise such as sequencing error.

Finally, pieces of evidence with an accompanying quality score of 10 or less (or in the case of SOLiD data, an average score of 10 or less for the two colors overlapping the evidence) were filtered out before smoothing the methylation data.

Smoothing via local likelihood estimation: Because the data was binomially distributed, we used local likelihood estimation[16]. This approach assumes that the p(L), the methylation level at genomic location L, is a smooth function of L; in other words, that CpGs that are close have similar methylation levels[17]. The local likelihood approach uses data within windows of predefined sizes to estimates p(L) and weighing data based on distance to L (based on a tricube kernel). In addition, the binomial model ensures that data points with high coverage receive greater weight. We defined two window sizes to detect the two different types of DMRs; for the blocks, a large window to detect low frequency differences, and a smaller window to detect high frequency differences, the small DMRs. For each sample, the smoothed data was evaluated on the same grid of data points termed "covered CpGs", consisting of those CpGs where at least two normals samples had coverage of at least 2.

For the small DMRs, the high-frequency analysis used a window size of 70 CpG or 1,000 bps; whichever generated a larger region. For the blocks, the low-frequency analysis used a window size of 500 CpGs or 2,000 basepairs; again whichever was larger. Note that the use of a tricube kernel ensures that data points far from the center of the window receive a smaller weight. This approach provided highly precise estimates of CpG methylation levels p(L) for each sample. The standard errors ranged from 0-0.11 (mean of 0.04) for the high frequency smoothing and 0.01-0.04 (mean 0.02) for low frequency. We obtained pair-wise correlations between methylation estimates for the three normal samples of 0.97, 0.96, and 0.96, and the three cancer samples of 0.87, 0.90, and 0.91, confirming coverage adequacy. An example of the results from the high-frequency smoothing is provided as Supplementary Figure 24. High-coverage data from selected regions confirm the highly accurate and precise estimates predicted by the statistical calculations as described in detail in the *Comparison of bisulfite capture and whole genome bisulfite sequencing Section* below

Accounting for biological variability: We then developed a method for finding differences based on t-statistics that take into account biological variability. We started with the highly precise estimates of $p_i(L)$ for each sample i at each CpG location L. We obtained the average difference between the three tumor samples and the three normal samples referred to as d(L). To properly account for biological variability (Supplementary Fig. 13) we estimated the standard error of d(L) using the normal samples. We used only the normal samples because as we demonstrated, with independent data, cancer samples are prone to high variability (Fig. 1) and here we are concerned only with DMRs In other words, we are not assuming that the cancer samples are biological replicates. The standard error se[d(L)] was therefore estimated as $\sigma(L)*\sqrt{(2/3)}$ with $\sigma(L)$ the standard deviation of the $p_i(L)$ for the three normal samples. To improve standard error estimates, we smoothed these using a running mean with a window size of 101 observations. To avoid inflated t-statistics as a result of artificially low variance, we set a threshold for the standard deviation of its 75th percentile, before computing the smoothed result. With the standard deviation in place we constructed the t-statistic $t(L) = d(L)/se[d(L)]$.

Correcting for low frequency effects: For the high frequency analysis the t-statistic was further corrected for low frequency changes. This allowed us to find local features, such as a hypermethylated

small DMR, inside global features, i.e. hypo- or hypermethylated block. We calculated this correction factor by forming a fixed grid of positions 2,000 bp apart in the genome, linearly interpolating the neighboring t-statistics to obtain measurements at these positions and then smoothing this dataset with a robust smoother based on the Huber family[16] and a bandwidth of 25,000 bp. It was important to use a fixed 2,000 bp grid instead of using the covered CpGs, otherwise we would correct out the local features we set out to identify. We then defined small DMRs as contiguous CpGs within 300 bp of each other, with the t-statistics above 4.6 or below -4.6 (corresponding to the 95th quantile of the empirical distribution of the t-statistics) and all differences in the same direction. For the low frequency analysis the t-statistics cutoff was 2 and contiguous CpG were defined as within 10,000 bps from each other.

Filtering and merging: These sets of regions formed our small DMRs and blocks that were subsequently filtered and processed according to the following criteria:

1) A small DMR needed to contain at least 3 covered CpGs and have at least 1 covered CpG per 300 bp. Furthermore, the mean difference in methylation percentages between tumors and normals across the small DMR had to be greater than 0.1.

2) A block needed to be longer than 5kb. Blocks containing CpG Islands with a mean methylation of less than 0.25 in the normal samples were separated into two. Putative blocks that were shorter than 5kb were included as small DMRs provided they satisfied the small DMR filters above.

After filtering, pairs of small DMRs were merged if they were less than 1kb apart, changed in the same direction (both hypermethylated or both hypomethylated), and had no covered CpGs in the area separating them. The final list of blocks is available as Supplementary Data 2, and of small DMRs as Supplementary Data 4.

The data from the adenoma samples were smoothed in the same way.

DMR Classification

Small DMRs were classified into categories based methylation profiles of the tumor and normal samples within the DMR and the two flanking regions (within 800bp). Based on these results, the DMRs that were discovered from data exploration could be classified into three types termed *loss of methylation boundaries*, *shifting of methylation boundaries*, and *novel hypomethylation* (Fig. 3). A mathematical algorithm was used to automatically classify DMRs. Briefly; mean methylation was computed for both tumor and normal samples within the DMR and in the flanking the DMR both upstream and downstream. This provided three numbers for each of the six samples. If all the normal samples showed high methylation values (>50%) in the flanking regions and low methylation values (<0.25%) and the tumor samples all showed intermediate values across DMR and flanking regions, the DMR was classified as *loss of methylation boundary.* If one of the flanking regions had low methylation values in both the normal and the tumor samples, the region was classified as a *shift of methylation boundary.* Finally, if all the normal samples showed high methylation values in the DMR and flanking regions while the tumor samples were lower in the DMR, the region was classified as *novel hypomethylation.* The details of the algorithm are best understood by viewing the computer code (made available upon request).

Comparison of bisulfite capture and whole genome bisulfite sequencing

The capture bisulfite experiment described above provided data for 474,829 CpGs (in one or more samples) from 39,262 regions. The genomic size of these regions ranged from 230 to 2,200 bp. For the analysis presented here we considered only the CpGs with coverage above 30x which resulted in 39,285, 107,332 and 86,855 CpGs in the normal samples and 125,611, 94,320, and 104,680 in the cancer. We computed an estimate of methylation for each CpG using simply the proportion of reads showing evidence of methylation for that CpG. We did not perform any averaging across genomic regions or

sample. We then compared the whole genome bisulfite sequencing data processed with BSmooth (referred to here as WGBS) to the high-coverage capture bisulfite (referred to as CAP).

The correlation between the WGBS and CAP data was 0.89, 0.89, and 0.87 for the normal samples and 0.83 for all three cancer samples. These results are quite remarkable given that the data were created at different times, by different experimental protocols, in different laboratories, and using different technology platforms. To determine how well the WGBS data corresponds to the CAP data spatially along the genome we visually inspected all genomic regions for which we had at least 50 CpGs with a coverage of at least 30x within a 5kb window. This yielded 49 such regions. Supplementary Figure 8 compares the result of BSmooth from WGBS with the single-base resolution methylation estimates from CAP for three such regions. The agreement between the smooth curves produced by BSmooth and the high-coverage CAP data is again remarkable.

Analysis of possible strand bias

Strand bias may be a problem in bisulfite sequencing data. Here we demonstrate that strand bias is not a concern with our data. We first noted that the sample specific proportion of reads originating from the Watson strand ranged from 0.503-0.504. We then examined each of the small DMRs and the blocks for possible strand bias by computing the percentage of the total evidence coming from the Watson strand for each of the reported regions. Specifically, we computed the sum of the Watson coverage of each CpG in the region divided by the sum of the coverage of each CpG and then averaged these numbers across the three normal samples and the three tumor samples separately. In total, 97.5% (tumors) and 97.4% (normals) of the small DMRs and 99.6% (tumors) and 99.8% (normal) of the blocks has a percentage of evidence coming from the Watson strand between 20% and 80%. Regions with extreme values of the percent evidence coming from the Watson strand were all very small and contained very few CpGs. For the convenience of users of our reported regions, columns with these statistics for each small DMR and block are included in Supplementary Tables 3, 7, 20, and 21.

Additional analyses

Pyrosequencing

To verify the accuracy of our methylation values obtained from BSmooth, we performed bisulfite pyrosequencing on the same 6 samples that were sequenced, for the small DMR regions shown in Figure 3. A 300 ng aliquot of genomic DNA from the sequenced samples were bisulfite converted and amplified using nested PCR (primers listed in Supplementary Data 13). The annealing temperature used for all PCR reactions was 50C. The resulting PCR products were used directly in pyrosequencing reactions, using an HS96A pyrosequencer (Qiagen). Plotting these loci shows good correspondence with our smoothed methylation values (Supplementary Fig. 9).

Defining Tissue-Specific Genes

We downloaded 529 gene expression microarrays from NCBI GEO representing 30 different tissues for which at least 5 biological replicates were available. The GEO accession numbers for these 529 microarrays are listed in Supplementary Table 23. We defined a tissue specific gene as a gene that was consistently expressed in 95% or more of the biological replicates for 5 or fewer tissues.

Sample-specific blocks and DMRs and their overlap

Sample specific blocks and DMRs were computed as per the outline above, by comparing a single tumor sample to all three normal samples. Each of the three sets of sample-specific blocks was found to be

highly concordant with the blocks obtained from the joint analysis of all the cancer and normal samples. Specifically, 95.1%, 98.3%, 96.9% of the bases covered by the three tumor-specific blocks overlap with the blocks from the joint analysis. Conversely 94.4%, 88.7%, 82.0% of the bases covered by the blocks from the joint analysis overlaps the three tumors specific blocks. All of these overlaps are highly significant (fisher's exact test, $p < 2.2e-16$). This demonstrated that the tumor-specific blocks are largely contained inside the blocks from the joint analysis. The list of sample-specific blocks and small DMRs can be found in Supplementary Table 20 and 21.

Co-occurrence of sample-specific blocks

To further investigate the extent to which sample-specific blocks co-occur, we analyzed the start and end positions of the blocks as follows. For each chromosome, we used the observed distribution of the sample-specific blocks to estimate the distance between block starts. For each chromosome, 1,000 simulated start positions of blocks were generated according to this distribution. We excluded chromosome Y due to the small number of sample-specific blocks. Each set of simulated start positions were constrained to the set of genomic CpG positions to take into account the fact that CpGs are not randomly distributed throughout the genome. We then picked one of the individuals to serve as a reference, and for each observed start position on the reference individual we computed the distance to the closest start site in each of the two other individuals. We also computed the distance between the reference individual and each of the 1,000 simulated sets of start positions. For illustration purposes, Supplementary Figure 10a shows the observed and simulated block start for a 20 Mb region of chromosome 1. In this analysis each of the three individuals was in turn used as a reference, yielding 6 sets of observed distances and 3,000 sets of expected (simulated) distances. The median distance for each of the 6 observed set of distances were smaller than the median of every single set of expected distances ($P<0.001$). We repeated this analysis for the end sites as well, obtaining the same results. Supplementary Figure 10b shows boxplots of the observed and expected distance distributions, where the observed distribution is the pool of all 6 observed distributions (the individual distributions are very similar) and the expected distribution is the pool of all 3,000 expected distributions (the individual distributions were again very similar).

Hypomethylation in blocks and repeat regions

Repeat regions were identified based on the UCSC repeatMasker track[18]. Based on the repeats and/or blocks, the genome was segmented into regions both repeats and blocks, repeats but not blocks, not repeats but blocks, and neither repeats nor blocks. The methylation levels were computed as the average of the high-frequency smoothed methylation levels of all CpGs in the 4 different regions. Density estimates were computed from the same distribution. Supplementary Table 3 describes the extent to which we were able to map CpGs inside repeat elements.

Enrichment of overlap between different genomic domains

For each pair of different genomic domains (like blocks and LOCKs) we form a 2x2 table containing the number of CpGs inside and outside the two genomic domains (like inside blocks and inside LOCKs, inside blocks and outside LOCKs, etc). Odds ratios and p-values were calculated using Fisher's exact test.

Copy number analysis

Estimates of copy number were based on the per-base coverage obtained after alignment. We did not apply the filters developed specifically for methylation measurements (described in the *Bisulfite alignment* Section). Note that the coverage we are considering here is not specific to CpGs: every genomic position is assigned a coverage value. We then computed, for each sample, the average coverage in non-overlapping 10,000bp windows, yielding two coverage vectors for each tumor-normal

pair denoted cov(T) and cov(N). For each tumor-normal pair we defined the corrected log-ratio: log2(CN) = log2(cov(T)) – log2(cov(N)) + c. Here c is a correction factor to account for different yields in each sequencing run; c is defined as the log of total sequencing yield of the normal sample divide by total yield of the tumor sample. The copy number log-ratios were segmented using circular binary segmentation (CBS)[19]. For illustrative purposes, copy number log-ratios and the associated segmentation on chromosome 20 were depicted (Supplementary Fig. 11a).

To determine if copy number had an effect on methylation estimates, each segment provided by CBS was divided into 100kb regions. For each of these regions we computed average copy number ratios as well as average methylation ratios. These were then plotted (Supplementary Fig. 11b) and no relationship between CNV and methylation blocks was observed.

Gene expression analysis

We obtained expression data from the gene expression barcode (rafalab.jhsph.edu/barcode). This resource combines all the expression data from the public repositories purportedly to standardize data in a way that allows one to call a gene expressed or not expressed [20]. From this source, we used two independent colon cancer datasets (Fig 5b: GSE8671[21] and Supplementary Fig 18: GSE4183[22,23]). To define hypervariable genes we performed an F-test using a across sample variance in tumor and normal samples computed from the original log expression. A gene was defined as expressed if it had a gene expression barcode standardized value above 2.54 (p=0.01). For the fibroblast analysis we downloaded datasets (GSE7890[24], GSE11418[25] , GSE11919[26]). These expression values from these datasets were also standardized using the gene expression barcode. The standardized values were used to determine if genes were expressed or not each sample.

To determine the correlation between small DMRs and expression, we considered a gene and a small DMR associated if the DMR was within 2,000 bps from the transcription start site of the gene; 6,869 genes mapped to a DMR in this way.

Note that because the focus of our paper is to report reproducible results about cancer, we have confirmed the inverse relationship between methylation and gene expression on completely independent sets (with multiple samples). But we also confirmed these results using one of the same samples for which we had methylation. Specifically, we obtained gene expression by hybridizing one of our normal/cancer pairs to an Affymetrix array (GEO accession number GSE13471). The inverse relationship was again confirmed (p<10^-15).

Gene expression variance analysis

Because the great majority of genes exhibit increased variance in cancer samples, standard statistical inference techniques do not guide the choice of a threshold to dichotomize genes by hypervariability. To demonstrate the indisputable association between hypomethylated blocks and hypervariabilty of gene expression we stratified genes by their across-sample standard deviation in cancer into 10 bins and for each bin we calculated the proportion of these genes that are in hypomethylated blocks. There is a clear (Supplementary Fig. 17), and statistically significant (p<0.01), direct relationship starting at about 20% and ending at 100%.

Gene ontology (GO) enrichment analysis

Throughout the text we described results from gene ontology (GO) enrichment analyses. These analyses were based on gene ontology enrichment analysis[27,28]. Specifically, for any given gene list we performed chi-squared test for association between genes in the list and GO categories. The analysis was carried out using the Bioconductor GOstats package[29].

Data Annotation

We obtained annotation from the UCSC genome browser based on hg19. In the cases where a data track was only available for hg18 or hg17, the UCSC liftOver tool[30]was used to map between builds of the human genome. Specifically we used the repeatMasker track[18], the RefSeq mRNA track[31], and the UCSC known genes track[32].

Laminin Associated Domain (LAD) coordinates were obtained from the NKI LADs track from UCSC, generated from microarrays in fibroblast cells[33]. PMDs were obtained from Lister et al., generated from bisulfite sequencing in fibroblast cells[12]. DNase I hypersensitive sites were obtained from the UCSC ENCODE track[34], using the H1es, Caco2rep1, and Caco2rep2, both the narrow and broad peak.

References

1. Irizarry, R.A. *et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* **41**, 178-86 (2009).
2. Bibikova, M. *et al.* High-throughput DNA methylation profiling using universal bead arrays. *Genome Res* **16**, 383-93 (2006).
3. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-93 (2003).
4. Leek, J.T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* **11**, 733-9 (2010).
5. McCullagh, P. & Nelder, J.A. *Generalized linear models*, xix, 511 p. (Chapman and Hall, London ; New York, 1989).
6. Aryee, M.J. *et al.* Accurate genome-scale percentage DNA methylation estimates from microarray data. *Biostatistics* **12**, 197-210 (2011).
7. Irizarry, R.A. *et al.* Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res* **18**, 780-90 (2008).
8. Venkatraman, E.S. & Olshen, A.B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657-663 (2007).
9. Bormann Chung, C.A. *et al.* Whole methylome analysis by ultra-deep sequencing using two-base encoding. *PLoS One* **5**, e9320 (2010).
10. Zon, G. *et al.* Formamide as a denaturant for bisulfite conversion of genomic DNA: Bisulfite sequencing of the GSTPi and RARbeta2 genes of 43 formalin-fixed paraffin-embedded prostate cancer specimens. *Anal Biochem* **392**, 117-25 (2009).
11. Deng, J. *et al.* Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat Biotechnol* **27**, 353-60 (2009).
12. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315-22 (2009).
13. Xi, Y. & Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* **10**, 232 (2009).
14. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-95 (2010).
15. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
16. Loader, C. *Local regression and likelihood*, (Springer Verlag, 1999).
17. Eckhardt, F. *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* **38**, 1378-85 (2006).

18. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* **16**, 418-20 (2000).
19. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-72 (2004).
20. Zilliox, M.J. & Irizarry, R.A. A gene expression bar code for microarray data. *Nat Methods* **4**, 911-3 (2007).
21. Sabates-Bellver, J. *et al.* Transcriptome profile of human colorectal adenomas. *Mol Cancer Res* **5**, 1263-75 (2007).
22. Gyorffy, B., Molnar, B., Lage, H., Szallasi, Z. & Eklund, A.C. Evaluation of microarray preprocessing algorithms based on concordance with RT-PCR in clinical samples. *PLoS One* **4**, e5645 (2009).
23. Galamb, O. *et al.* Reversal of gene expression changes in the colorectal normal-adenoma pathway by NS398 selective COX2 inhibitor. *Br J Cancer* **102**, 765-73 (2010).
24. Smith, J.C., Boone, B.E., Opalenik, S.R., Williams, S.M. & Russell, S.B. Gene profiling of keloid fibroblasts shows altered expression in multiple fibrosis-associated pathways. *J Invest Dermatol* **128**, 1298-310 (2008).
25. Chen, Y. *et al.* Developing and applying a gene functional association network for anti-angiogenic kinase inhibitor activity assessment in an angiogenesis co-culture model. *BMC Genomics* **9**, 264 (2008).
26. Duarte, T.L., Cooke, M.S. & Jones, G.D. Gene expression profiling reveals new protective roles for vitamin C in human skin cells. *Free Radic Biol Med* **46**, 78-87 (2009).
27. Dennis, G., Jr. *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4**, P3 (2003).
28. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).
29. Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257-8 (2007).
30. Fujita, P.A. *et al.* The UCSC Genome Browser database: update 2011. *Nucleic Acids Res* (2010).
31. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61-5 (2007).
32. Hsu, F. *et al.* The UCSC Known Genes. *Bioinformatics* **22**, 1036-46 (2006).
33. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948-51 (2008).
34. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).

	Total genome coverage after filtering	Total # Cs, Ts aligning to CpG cytosines after filtering	Average coverage of mapable CpG cytosines
Normal 1	13,213,005,428	116,394,219	5.578
Cancer 1	13,106,826,609	118,994,714	5.703
Normal 2	13,137,358,383	119,303,310	5.717
Cancer 2	13,411,947,895	119,273,248	5.716
Normal 3	12,592,027,592	113,220,232	5.426
Cancer 3	12,593,520,085	113,929,959	5.460
Adenoma 1	10,965,839,892	102,341,491	4.905
Adenoma 2	11,017,476,026	95,652,482	4.584

Supplementary Table 1: Whole-genome bisulfite SOLiD sequencing coverage both genome-wide and for CpG cytosines. Average coverage of mappable CpG cytosines (column 3) was calculated by dividing the total number of pieces of C and T nucleotide evidence aligning to CpG cytosines after filtering (column 2) by the total number of CpG dinucleotides in GRCh37 that are “mappable” in our experiment: 20.9 million. A CpG was considered “mappable” if it was overlapped by at least one non-ambiguous reference 50 bp substring; non-ambiguous substrings are those for which no other 50 bp reference substring exists within 3 mismatches.

	Total nucleotide coverage after filtering	Total # Cs, Ts aligning to CpG cytosines after filtering	Average coverage of covered CpG cytosines	Number of CpGs with >= 30x coverage
Normal 1	200,586,597	5,574,845	22.3	39,285
Cancer 1	1,105,875,090	31,003,296	106.4	107,332
Normal 2	616,177,235	14,730,151	46.1	86,855
Cancer 2	1,137,330,587	29,033,239	82.3	125,611
Normal 3	834,907,436	18,480,903	67.4	94,320
Cancer 3	1,081,926,255	24,533,799	81.2	104,680

Supplementary Table 2: Capture bisulfite Illumina GA II sequencing. In total, 474,829 CpGs were covered by at least one read in at least one sample.

Repeat Family	Genomic size (bp)	Number of CpGs	Percent mappable CpGs	Percent covered CpGs
acro	31082	606	12.4	27.1
Alu	307842860	7129208	36.1	18
centr	8244270	81346	53.7	57.8
CR1	10918988	60817	96.8	93.8
Deu	180434	1386	99	97.8
DNA	341642	1968	94.8	91.6
DNA?	273824	1635	99.6	96.3
Dong-R4	121003	635	99.7	96.1
ERV	192131	1255	95.5	90.4
ERV1	83536316	857956	62.8	55.7
ERVK	8845044	118035	35.6	31.6
ERVL	56261410	376293	90	87.2
ERVL?	418361	2550	96.2	95.1
ERVL-MaLR	111131119	741120	85.4	79.8
Gypsy	2312231	13904	96.7	93.9
Gypsy?	1467092	8972	98.3	95.4
hAT	1686883	10028	95.8	88.2
hAT?	505238	3093	98	93.6
hAT-Blackjack	3436024	21096	97.1	91.9
hAT-Charlie	45149528	314072	93.9	87.5
hAT-Tip100	6635588	44491	96.6	91.9
Helitron	388235	2219	97.3	92.6
Helitron?	66137	326	99.4	92.9
L1	512818213	2714809	64	56.5
L1?	6868	39	100	100
L2	104388187	689060	96.9	94.2
Low_complexity	17235743	202393	92.5	63.4
LTR	476961	2548	94.5	92.6
LTR?	21980	141	100	97.9

Merlin	17762	130	91.5	88.5
MIR	84839808	572596	97	95.1
MuDR	692052	4514	94.1	88.6
Other	4015047	205878	6.9	2.3
Penelope?	10499	58	100	94.8
PiggyBac	500519	4325	84	70.9
PiggyBac?	44319	305	99	95.1
RNA	119461	1653	81.1	74.5
rRNA	176927	4124	81.5	77.6
RTE	3661238	18331	96.9	92.6
RTE-BovB	75161	445	98.2	97.3
Satellite	4014235	60731	48.9	55.3
scRNA	123332	989	89.4	73.8
Simple_repeat	26384021	225215	85.6	54.3
SINE	162234	1209	99.5	97.7
SINE?	45384	339	97.1	94.7
snRNA	341832	5166	87.9	74.7
srpRNA	264574	3318	82.8	73.2
TcMar	320139	1539	96.7	89.3
TcMar?	628300	4179	96.8	92.8
TcMar-Mariner	2839480	21056	87.4	71.9
TcMar-Tc2	1676142	9019	95.8	88.8
TcMar-Tigger	34040312	215118	93.3	86.1
telo	254535	10983	31.8	22.8
tRNA	337404	5403	84.9	72.1
Unknown	1281847	7534	97.7	95.1
Unknown?	18418	172	98.8	97.7
All repeats	**1450353676**	**14777741**	**56.2**	**43.8**

Supplementary Table 4: Mappability of repeat families. For each repeat family in the UCSC repeatMasker track, we computed the size of the repeat family (in bp and in number of CpGs), its mappability and its coverage in percent CpGs. Note that the repeat family names ending in “?” designate repeats which the UCSC repeatMasker track defines as questionable.

Genomic domain	Size (in GB)	Size (in millions of CpGs)	Overlap with blocks (in GB)	Overlap with blocks (in millions of CpGs)	Odds Ratio
Repeats	1.45	14.8	1.04GB	9.33	1.4
PMDs	1.23	10.0	1.14GB	8.45	6.5
LOCKs	0.77	5.8	697MB	5.06	6.8
LADs	1.14	8.6	989MB	7.10	4.9

Supplementary Table 5: The overlap of various genomic domains with differentially methylated regions in colon cancer. The size of genomic domains are shown in column 1 in gigabases and column 2 in number of CpGs. Column 3 shows the overlap in gigabases and Column 4 in number of CpGs. Column 5 shows the observed to expected by chance odds ratio. All the overlaps were statistically significant ($p < 2.2 \times 10^{-16}$).

	Hypervariable CpGs in all tissue types (custom array)		Non-variably methylated CpGs in colon cancer (Illumina 27k)
	Hypomethylated	**Hypermethylated**	
Number of CpGs	81	52	16,049
Inside Hypomethylated Blocks	63%	4%	13%
Inside Hypermethylated Blocks	5%	37%	2%

Supplementary Table 6. Enrichment of hypervariable methylated loci identified by the custom Illumina array in blocks identified by bisulfite sequencing. We divided the 157 CpGs that showed a statistically significant variation increase in all five tissues assayed with our custom array into hypermethylated in cancer and hypomethylated in cancer. The first two columns show that a high percentage of these hypervariable methylated CpGs are in blocks identified by colon cancer/normal bisulfite sequencing. The hypervariable methylated loci show a consistent direction of methylation change with the bisulfite sequencing result. To perform an enrichment analysis, we combined the loci in colon cancer identified with our custom array with those identified in the Illumina HumanMethylation27 array. There is a significant enrichment of hypomethylated hypervariable loci in hypomethylated blocks identified by sequencing (P-value < $1x10^{-16}$, Fisher test).

	Methylation status in normals	Total	Hypo	No change	Hyper
All islands	Unmethylated (<= 0.2)	16184	0.1%	83.2%	16.7%
	Partial methylated (>= 0.2, <=0.8)	4796	17.0%	46.7%	36.3%
	Methylated (>= 0.8)	5527	24.0%	75.9%	0.1%
Promoters	Unmethylated (<= 0.2)	11050	0.0%	88.5%	11.5%
	Partial methylated (>= 0.2, <=0.8)	1007	6.6%	50.1%	43.3%
	Methylated (>= 0.8)	231	22.1%	77.9%	0.0%
Genic	Unmethylated (<= 0.2)	13030	0.1%	85.6%	14.4%
	Partial methylated (>= 0.2, <=0.8)	2950	15.8%	46.6%	37.6%
	Methylated (>= 0.8)	4295	18.4%	81.6%	0.1%
Intergenic	Unmethylated (<= 0.2)	1633	0.2%	70.1%	29.7%
	Partial methylated (>= 0.2, <=0.8)	1463	21.5%	47.4%	31.0%
	Methylated (>= 0.8)	1137	45.5%	54.4%	0.2%
Repeats	Unmethylated (<= 0.2)	7386	0.0%	83.0%	17.0%
	Partial methylated (>= 0.2, <=0.8)	928	5.3%	47.1%	47.6%
	Methylated (>= 0.8)	316	12.7%	87.3%	0.0%

Supplementary Table 8: Methylation values observed in CpG islands in cancer compared to normal samples, stratified using location relative to known genes. This is a subdivision of Table 2. For this table, a CpG island may belong to more than one category (overlaps promoter region, overlaps genic including introns region, overlaps intergenic region, and overlaps repeat region). Average methylation values in each island were averaged across subject for cancer and normal samples separately. Note that in normal samples promoter CpG Islands are largely unmethylated, and of these ~12% become methylated in cancer. In contrast intergenic islands show a more balanced proportion of methylated and unmethylated state in normal tissue (35% versus 26%) with 45% hypomethylated in cancer.

	Associated Genes	Inversely correlated genes	Percent Inverse correlated
Shift of boundary (hypomethylated)	2,273	1,192	52%
Novel hypomethylation	38	17	45%
Other hypomethylated	442	192	43%
Shift of boundary (hypermethylation)	1,893	532	30%
Loss of boundary (hypermethylation)	1,119	346	31%
Other hypermethylated	1,204	335	28%

Supplementary Table 9: Gene expression negatively correlates with methylation in small DMRs. We mapped each gene represented in a microarray experiment to the closest small DMR, with a gene and a DMR considered associated if the DMR was within 2kbp from the transcription start site of the gene; 6,869 genes were mapped and are represented in the table. For each of the small DMR classes, as defined in main text, we computed the number of associated genes that were differentially expressed (FDR<0.05) and had an inverse relationship.

Gene ontology term	Expected count	Count	Size	Odds ratio	P-value	Q-value
Mitotic cell cycle	11.6	31	185	3.1	3.8×10^{-7}	0.00036
Cell cycle process	18.7	42	297	2.6	5.4×10^{-7}	0.00036
Mitosis	9.9	25	157	2.9	1.5×10^{-5}	0.006
Positive regulation of ubiquitin-protein ligase activity during mitotic cell cycle	3.9	14	61	4.5	2.0×10^{-5}	0.0071
Organelle fission	13.7	30	216	2.5	3.6×10^{-5}	0.0097

Supplementary Table 10: Gene ontology enrichment analysis of differentially expressed genes associated with hypomethylated boundary shifts. The differentially expressed genes comparing colon cancer samples to normal samples were divided into two groups: those associated hypomethylated boundary shifts and the rest. The table shows the five categories that with statistically significant enrichment (FDR<0.01).

Term	Exp Count	Count	Size	Gene IDs	Odds Ratio	P-value	Q-value
collagen catabolic process	0.160	6	20	MMP10; MMP7; MMP3; KLK6; MMP19; MMP1	56.3	8.1e-09	0.0000021
multicellular organismal macromolecule metabolic process	0.288	7	36	MMP10; MMP7; MMP3; KLK6; IL6; MMP19; MMP1	32.0	1.2e-08	0.0000021
biopolymer catabolic process	3.486	15	425	MMP10; MMP7; MMP3; TMPRSS3; FAP; KLK6; ADAMDEC1; KLK12; TMPRSS3; MMP19; KLK8; REN; HGF; MEP1A; KLK12; PCSK1; MMP1	5.0	1.8e-06	0.0001989
catabolic process	3.565	15	438	MMP10; MMP7; MMP3; TMPRSS3; FAP; KLK6; ADAMDEC1; KLK12; TMPRSS3; MMP19; KLK8; REN; HGF; MEP1A; KLK12; PCSK1; MMP1	4.9	2.3e-06	0.0001989
inflammatory response	1.999	11	251	CXCL11; CCL26; IL1A; S100A12; IL8RB; SERPINA3; CHST4; IL17A; REG3A; IL22; C4BPA	6.3	4.8e-06	0.0003384
protein metabolic process	2.697	12	342	MMP10; MMP7; MMP3; TMPRSS3; FAP; ADAMDEC1; KLK12; TMPRSS3; MMP19; KLK8; MEP1A; INHBA; KLK12; MMP1	5.2	1.4e-05	0.0007995
cell-cell signaling	2.294	11	286	STC1; CXCL11; STC1; CCL26; WISP3; CHST4; IL6; IL17A; STC1; HGF; IL22; INHBA; PCSK1	5.4	1.8e-05	0.0008855
acute-phase response	0.206	4	26	SERPINA3; IL6; REG3A; IL22	23.6	4.9e-05	0.0021309
ectoderm development	0.117	3	15	KRT6A; KRT6B; ELF5	32.8	1.9e-04	0.0075583

Supplementary Table 13: Gene ontology enrichment analysis of hypervariable genes associated with blocks. Genes represented in the microarray were divided into two groups: genes contained in hypomethylated blocks showing hypervariability in cancer and the rest. The table shows the nine categories with statistically significant enrichment (FDR<0.01).

	Reads	Unique	Unaligned	Non-unique	Uniquely aligned to λ phage	Conversion percent (λ estimate)
Normal 1 Flowcell 1	485,990,920	183,844,653	226,160,599	75,985,667	6,762,031	99.7789
Normal 1 Flowcell 2	491,108,959	180,074,869	235,598,979	75,435,110	6,974,109	99.7807
Cancer 1 Flowcell 1	495,809,693	185,828,257	233,836,993	76,144,442	8,526,067	99.7456
Cancer 1 Flowcell 2	482,952,465	178,456,624	230,494,869	74,000,971	8,122,750	99.7472
Normal 2 Flowcell 1	496,397,317	180,791,619	237,818,636	77,787,061	7,270,350	99.7646
Normal 2 Flowcell 2	503,561,286	182,437,294	242,263,685	78,860,306	7,538,526	99.7657
Cancer 2 Flowcell 1	497,625,604	187,406,059	232,909,571	77,309,973	8,251,881	99.7804
Cancer 2 Flowcell 2	494,805,108	184,876,404	233,935,851	75,992,852	8,261,281	99.7860
Normal 3 Flowcell 1	489,558,995	168,584,119	251,409,912	69,564,963	8,817,079	99.7953
Normal 3 Flowcell 2	495,129,950	185,140,655	234,292,582	75,696,712	9,550,241	99.7937
Cancer 3 Flowcell 1	491,209,978	170,520,921	249,977,651	70,711,405	7,104,618	99.7919
Cancer 3 Flowcell 2	475,735,953	179,289,913	222,767,391	73,678,648	7,295,636	99.7930
Adenoma 1 Flowcell 1	477,819,340	153,546,992	259,594,989	64,677,358	3,457,741	99.7684
Adenoma 1 Flowcell 2	474,211,342	149,570,667	260,639,196	64,001,478	3,337,933	99.7678
Adenoma 2 Flowcell 1	479,069,565	165,738,601	242,575,131	70,755,832	12,991,935	99.7021
Adenoma 2 Flowcell 2	458,864,972	150,866,083	244,060,513	63,938,375	10,969,639	99.7055
Total	**7,789,851,447**	**2,786,973,730**	**3,838,336,548**	**1,164,541,153**	**125,231,817**	**N/A**
Average	**486,865,715**	**174,185,858**	**239,896,034**	**72,783,822**	**7,826,989**	**99.7667**

Supplementary Table 14: Sequencing and alignment results for the 7.79 billion bisulfite reads obtained from 8 runs (16 flowcells) of a SOLiD 3+ instrument. Alignment was performed against a collection of reference sequences consisting of the GRCh37 human genome assembly, including mitochondrial DNA and "unplaced" contigs, plus the sequence of the spiked-in λ DNA. A read is said to align "uniquely" if it has exactly one valid alignment to the reference according to the alignment policy. A read is said to align "non-uniquely" if it has more than one valid alignment according to the alignment policy. A read "fails" to align if it has zero valid alignments. Cytosine conversion percentage is estimated as the fraction of high-quality evidence from unique λ phage alignments indicating lack of methylation.

	Reads	Unique	Failed to align	Non-unique
Normal 1	3,471,782	2,548,012	405,480	518,262
Cancer 1	18,054,443	14,048,143	1,539,281	2,466,307
Normal 2	9,927,197	7,826,766	964,752	1,135,416
Cancer 2	18,132,540	14,444,610	1,595,357	2,091,814
Normal 3	12,939,010	10,599,360	933,981	1,405,235
Cancer 3	16,821,937	13,736,162	1,238,826	1,846,244
Total	**79,346,909**	**63,203,053**	**6,677,677**	**9,463,278**
Average	**13,224,002**	**10,533,842**	**1,112,946**	**1,577,213**

Supplementary Table 15: Sequencing and alignment results for 79.3 million capture bisulfite reads obtained from Illumina GA II instrument. Alignment was performed against a collection of reference sequences consisting of the GRCh37 human genome assembly, including mitochondrial DNA and "unplaced" contigs. A read is said to align "uniquely" if it has exactly one valid alignment to the reference according to the alignment policy. A read is said to align "non-uniquely" if it has more than one valid alignment according to the alignment policy. A read "fails" to align if it has zero valid alignments.

	Reads providing human CpG evidence	Raw pieces of human CpG evidence	Filtered pieces of human CpG evidence	Filtered evidence indicating presence of methylation	Filtered evidence indicating lack of methylation	% CpGs covered by ≥1 piece of filtered evidence
Normal 1 Flowcell 1	54,178,356	74,924,756	68,839,724	48,253,347	20,586,377	65.158%
Normal 1 Flowcell 2	53,689,203	74,623,785	68,199,772	47,773,815	20,425,957	65.045%
Cancer 1 Flowcell 1	55,834,653	77,831,608	71,498,083	45,243,596	26,254,487	65.103%
Cancer 1 Flowcell 2	53,902,278	75,253,470	68,697,400	43,321,299	25,376,101	64.479%
Normal 2 Flowcell 1	54,492,135	76,436,939	70,022,135	49,881,756	20,140,379	65.797%
Normal 2 Flowcell 2	55,012,691	77,239,693	70,641,796	50,296,004	20,345,792	65.978%
Cancer 2 Flowcell 1	55,815,773	77,542,380	71,153,086	42,104,541	29,048,545	65.526%
Cancer 2 Flowcell 2	54,867,691	76,121,856	69,486,614	40,787,828	28,698,786	65.271%
Normal 3 Flowcell 1	50,056,580	70,114,583	63,620,050	43,447,381	20,172,669	64.004%
Normal 3 Flowcell 2	54,385,230	75,777,367	69,162,172	47,580,371	21,581,801	65.333%
Cancer 3 Flowcell 1	51,426,265	71,942,006	65,509,795	35,741,629	29,768,166	63.954%
Cancer 3 Flowcell 2	53,527,518	74,427,549	68,200,974	37,298,457	30,902,517	64.356%
Adenoma 1 Flowcell 1	47,841,436	67,596,302	61,069,078	40,572,495	20,496,583	63.580%
Adenoma 1 Flowcell 2	46,522,581	65,753,563	59,413,620	39,519,156	19,894,464	63.125%
Adenoma 2 Flowcell 1	46,102,652	63,325,754	58,045,004	33,949,586	24,095,418	61.767%
Adenoma 2 Flowcell 2	43,224,022	59,918,936	54,862,907	32,006,566	22,856,341	60.444%
Total	**830,879,064**	**1,158,830,547**	**1,058,422,210**	**677,777,827**	**380,644,383**	**N/A**
Average	**51,929,942**	**72,426,909**	**66,151,388**	**42,361,114**	**23,790,274**	**64.308%**

Supplementary Table 16: Methylation evidence results for the 831 million bisulfite reads that aligned uniquely to the GRCh37 human genome assembly and overlapped at least one CpG cytosine. A piece of CpG evidence occurs when an alignment overlaps the cytosine position of a CpG in the reference sequence and the overlapping alignment nucleotide is either a T (indicating a lack of methylation) or a C (indicating presence of methylation). A filter is applied to remove nucleotide evidence that (a) is refuted by one or both of the overlapping colors from the original read, or (b) is within 4 positions of either end of the nucleotide alignment.

	Reads providing human CpG evidence	Raw pieces of human CpG evidence	Filtered pieces of human CpG evidence	Filtered evidence indicating presence of methylation	Filtered evidence indicating lack of methylation
Normal 1	2,187,604	6,759,584	6,759,584	2,922,481	3,837,103
Cancer 1	12,101,267	37,662,368	37,662,368	18,012,291	19,650,077
Normal 2	6,486,766	17,935,202	17,935,202	7,858,465	10,076,737
Cancer 2	12,139,505	35,156,131	35,156,131	15,644,426	19,511,705
Normal 3	8,607,882	22,477,332	22,477,332	9,783,575	12,693,757
Cancer 3	11,236,160	29,853,287	29,853,287	12,443,798	17,409,489
Total	**52,759,184**	**149,843,904**	**149,843,904**	**66,665,036**	**83,178,868**
Average	**8,793,197**	**24,973,984**	**24,973,984**	**11,110,839**	**13,863,145**

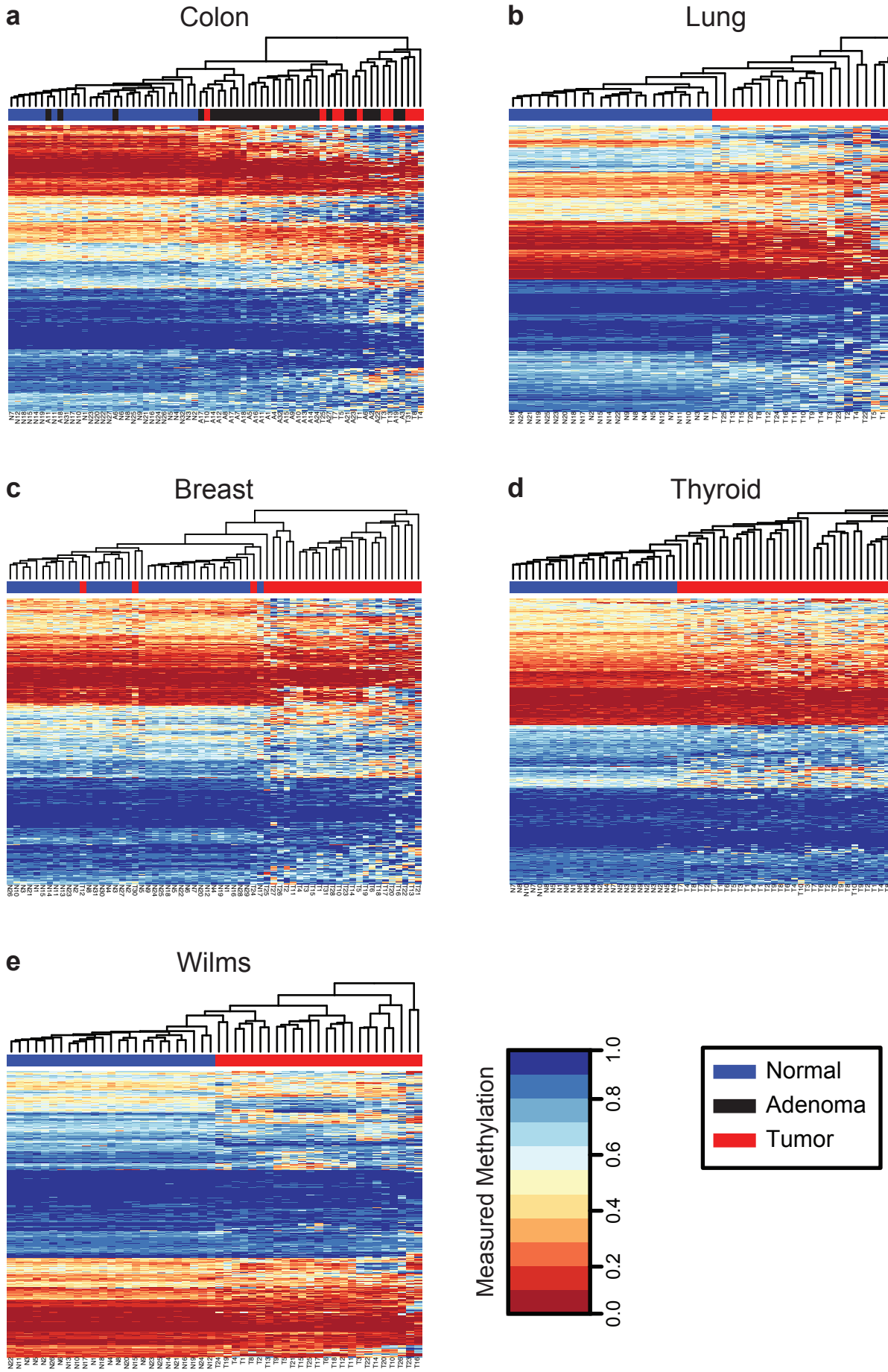
Supplementary Table 17: Methylation evidence results for the 63.2 million Illumina GAII capture bisulfite reads that aligned uniquely to the GRCh37 human genome assembly and overlapped at least one CpG cytosine. A piece of CpG evidence occurs when an alignment overlaps the cytosine position of a CpG in the reference sequence and the overlapping alignment nucleotide is either a T (indicating a lack of methylation) or a C (indicating presence of methylation). A filter is applied to remove nucleotide evidence that (a) is refuted by one or both of the overlapping colors from the original read, or (b) is within 4 positions of either end of the nucleotide alignment.

	Percent of CpGs in human genome covered with filtered evidence at depth ≥ threshold							
Threshold	**Normal 1**	**Cancer 1**	**Normal 2**	**Cancer 2**	**Normal 3**	**Cancer 3**	**Adenoma 1**	**Adenoma 2**
1	73.395%	73.279%	73.934%	73.636%	73.555%	72.994%	73.062%	71.459%
2	64.399%	63.918%	65.169%	64.567%	63.869%	63.247%	62.381%	60.906%
3	56.813%	56.119%	57.880%	57.091%	55.734%	55.099%	53.244%	51.613%
4	49.137%	48.461%	50.469%	49.640%	47.684%	47.116%	44.372%	42.406%
5	41.348%	40.917%	42.832%	42.148%	39.706%	39.335%	35.876%	33.569%
6	33.796%	33.753%	35.286%	34.872%	32.115%	32.000%	28.103%	25.573%
7	26.804%	27.219%	28.188%	28.088%	25.232%	25.391%	21.338%	18.770%
8	20.635%	21.496%	21.847%	22.042%	19.270%	19.674%	15.731%	13.296%
9	15.442%	16.658%	16.446%	16.862%	14.321%	14.917%	11.282%	9.114%
10	11.254%	12.693%	12.040%	12.596%	10.369%	11.087%	7.892%	6.061%

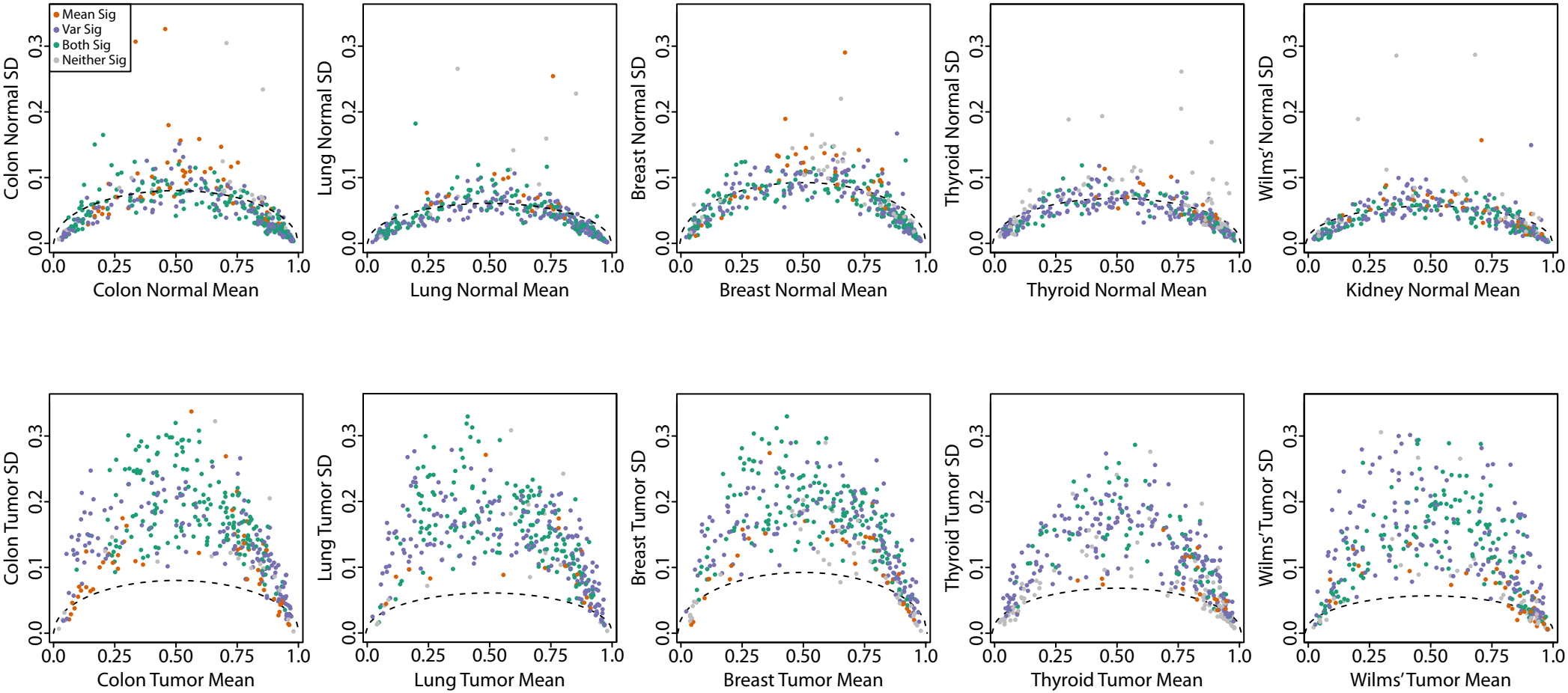
Supplementary Table 18: Fraction of CpGs in the GRCh37 human genome assembly covered by at least one through ten pieces of filtered evidence from whole-genome bisulfite SOLiD data. Each sample is sequenced on two flowcells of a SOLiD 3+ instrument; here, results are calculated after pooling evidence from the two flowcells.

Sample	CpG positions: Percent T	CpG positions: Percent C	Non-CpG C positions: Percent T	Non-CpG C positions: Percent C	Estimated % cytosines unconverted
Normal 1	30.03%	69.84%	99.76%	0.20%	0.22%
Cancer 1	36.95%	62.92%	99.73%	0.23%	0.25%
Normal 2	28.87%	70.99%	99.74%	0.22%	0.24%
Cancer 2	41.16%	58.71%	99.77%	0.20%	0.22%
Normal 3	31.48%	68.39%	99.77%	0.20%	0.21%
Cancer 3	45.39%	54.48%	99.77%	0.19%	0.21%
Adenoma 1	33.57%	66.32%	99.73%	0.24%	0.23%
Adenoma 2	41.68%	58.17%	99.70%	0.25%	0.30%

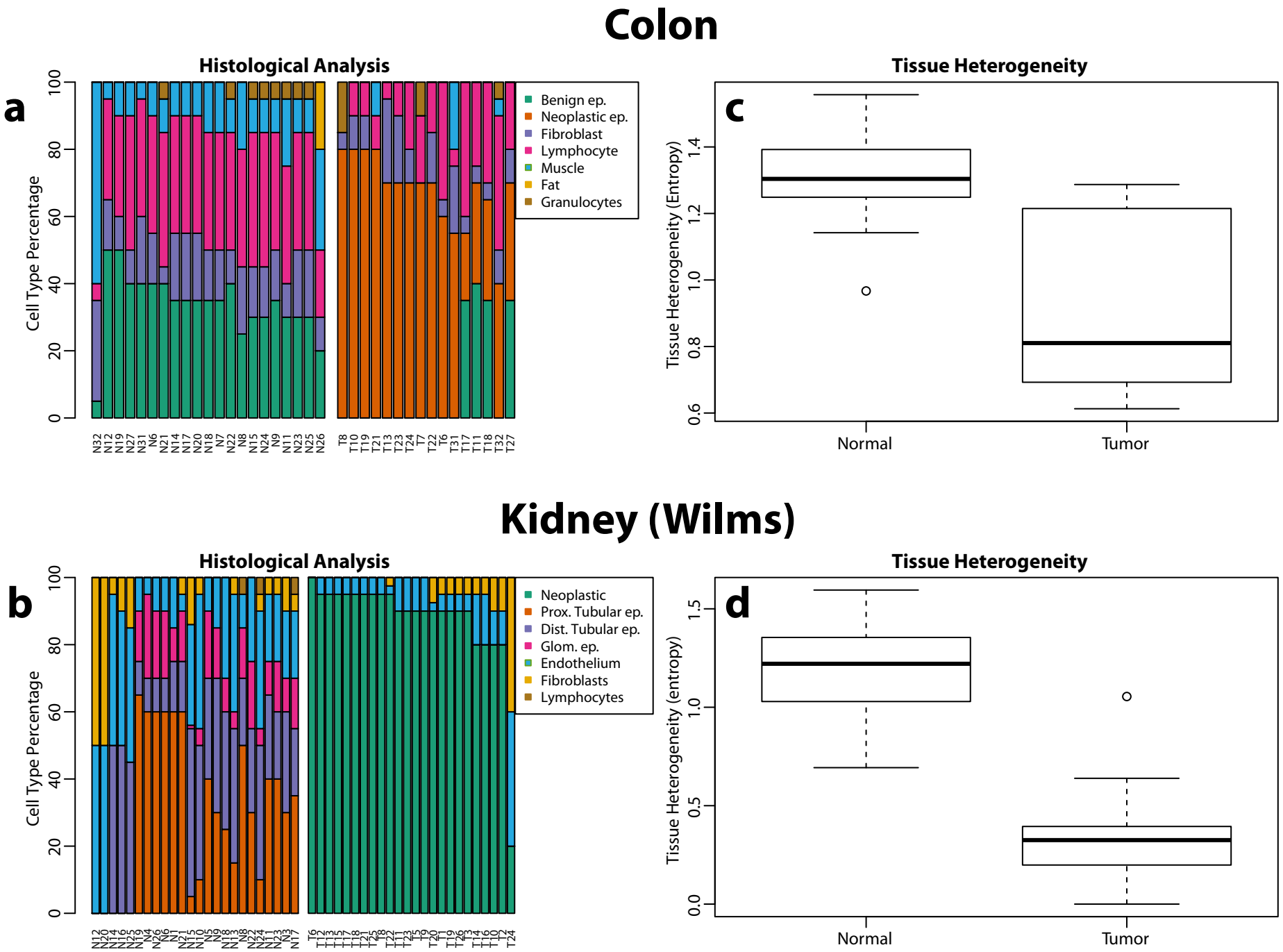
Supplementary Table 19: Fraction of filtered nucleotide evidence from whole-genome bisulfite SOLiD data where evidence indicates presence of a T or C. Filtered nucleotide evidence consists of evidence (a) from reads that aligned uniquely, (b) where both overlapping colors from the original read agree with the decoded nucleotide and, (c) where nucleotides within 4 positions of either end of the alignment are excluded. The first two columns show the global fractions of Ts and Cs covering CpG cytosines. The third and fourth columns show the global fractions of Ts and Cs covering non-CpG cytosines. For comparison, the final column shows the unconverted cytosine rate (one minus the conversion rate) estimated from the λ phage alignments.



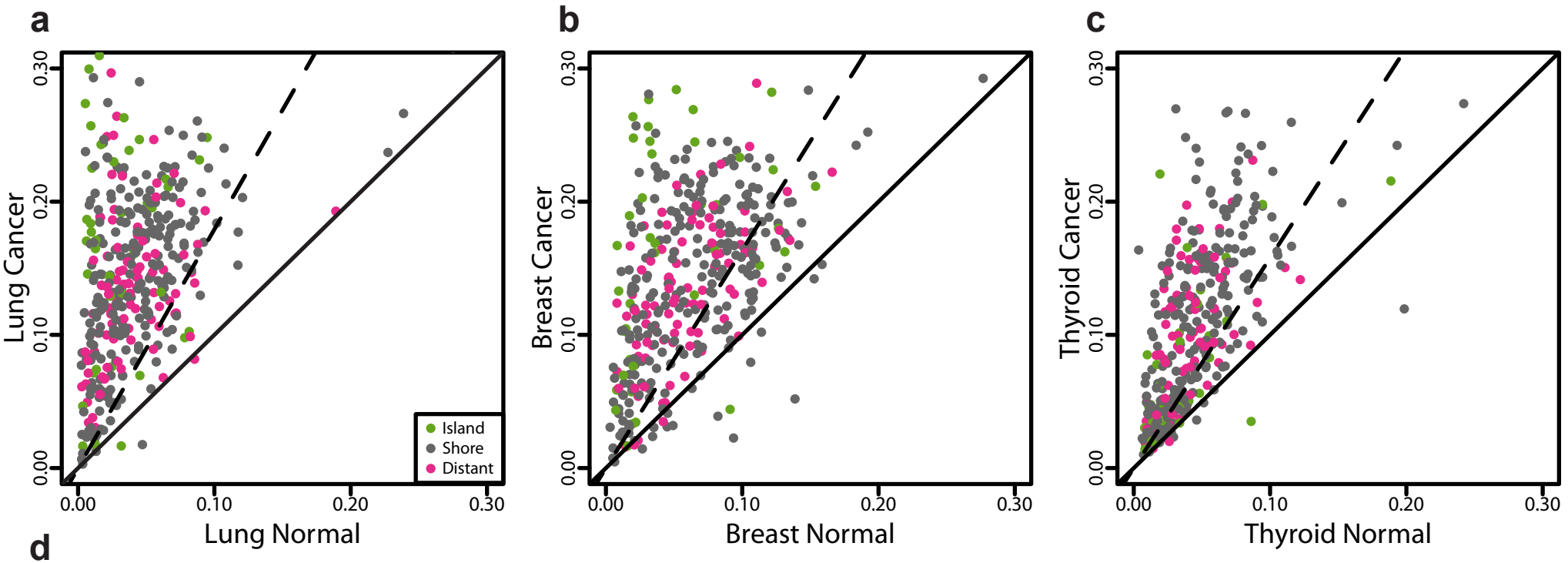
Supplementary Figure 1: Methylation levels of differentially methylated CpGs in colon cancer largely differentiate cancer from normal in colon(a), lung(b), breast(c), thyroid(d) and kidney(e)(Wilms) tissues. Columns and rows in each panel are ordered by a hierearchical clustering of methylation profiles using Euclidean distance. The heights of dendrogram branches, larger between tumor samples than between normal samples, illustrate the increased across-sample variability in cancer seen in a majority of CpGs in all tissues.

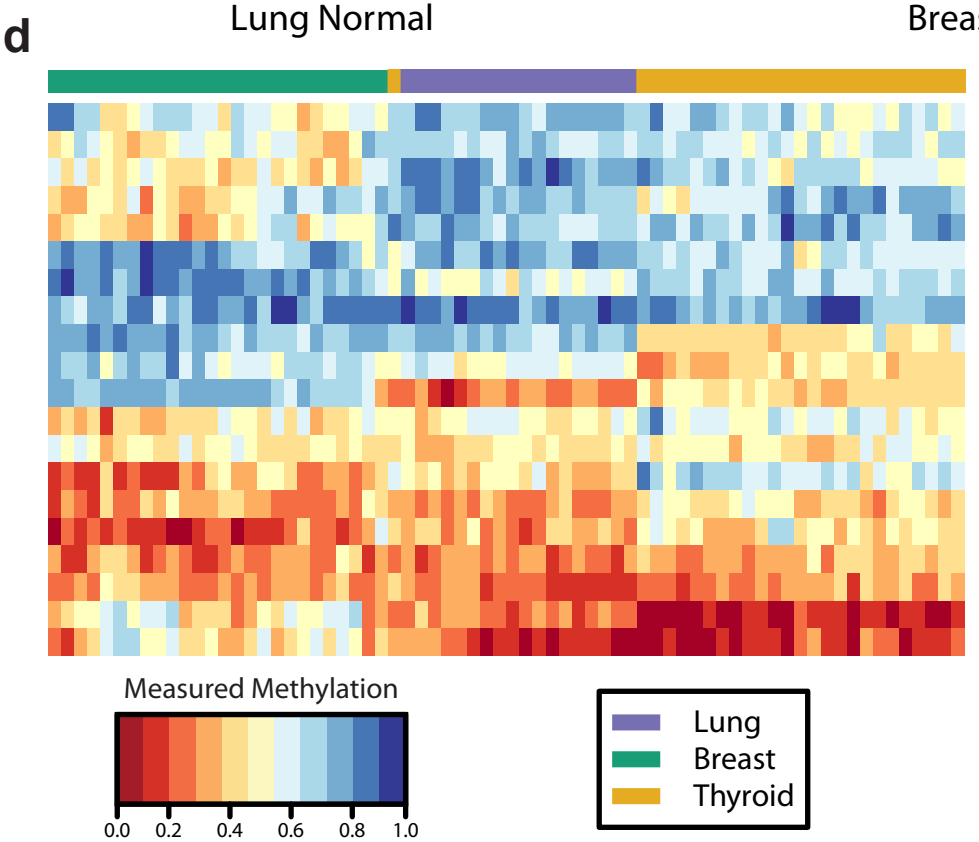


Supplementary Figure 2: Increased variability is not due to changes in mean methylation levels. Each column plots mean versus standard deviation of methylation values for normal and cancer in colon, lung, breast, thyroid and kidney (Wilms' tumor). The dotted line indicates the expected variance from the binomial model at each mean methylation level. Increased variability is clearly observed in cancer along the range of methylation values. CpGs are coded to indicate significant differences in mean only (orange), variance only (purple), both (green), or neither(grey).

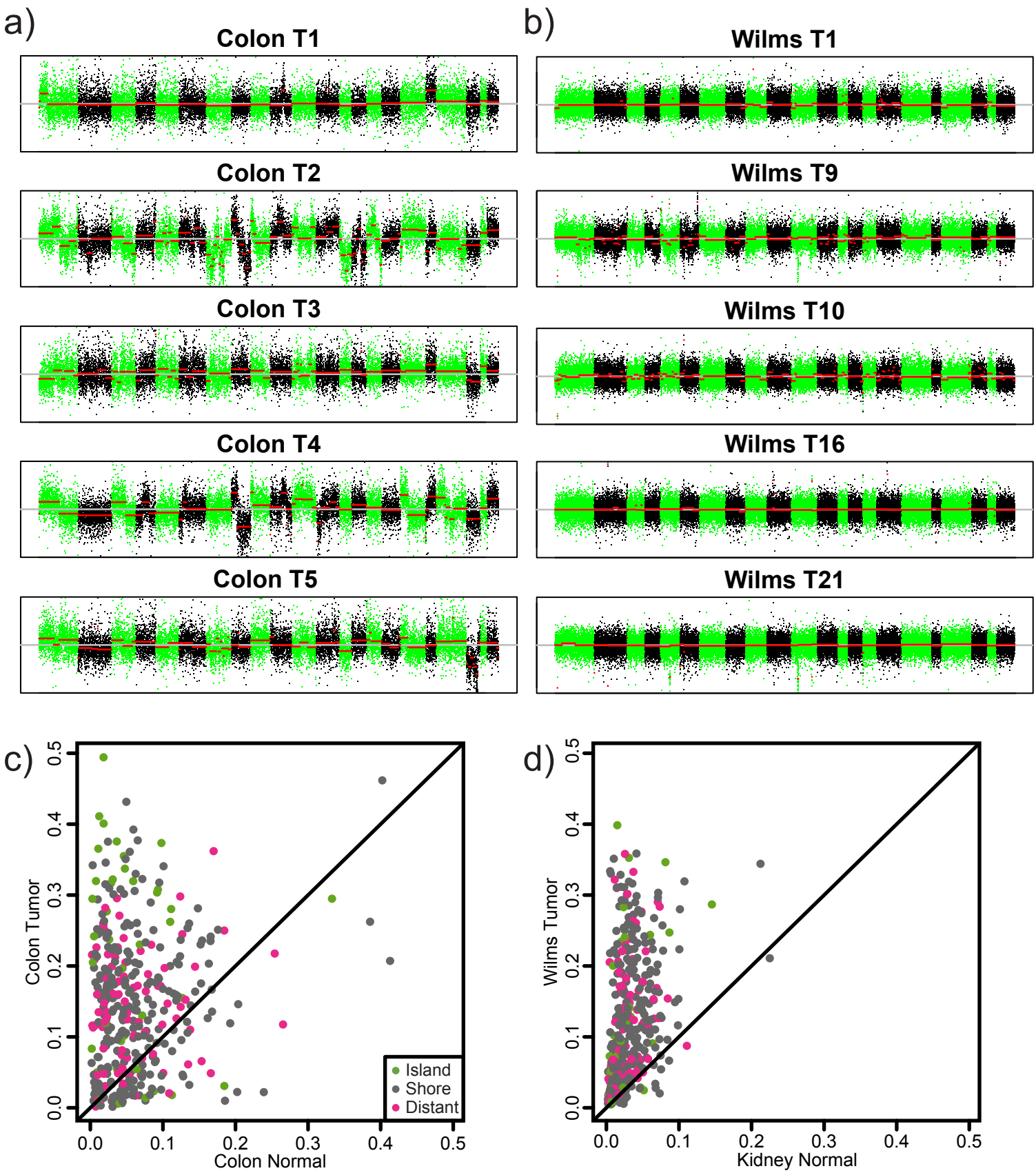


Supplementary Figure 3: Histological analysis of colon samples shows increased methylation variation is not due to cellular heterogeneity of samples. (a) and **(b)** Summary of the histological analysis. Each bar corresponds to a sample ((a) colon or (b) kidney) with normal samples denoted N and tumors as T. The tumor samples are actually more homogenous since they are primarily composed of neoplastic epithelial cells. **(c)** and **(d)** Quantification of sample heterogeneity using the entropy of the cellular composition of each sample. On average normal samples are more heterogeneous than tumor samples in cellular composition.

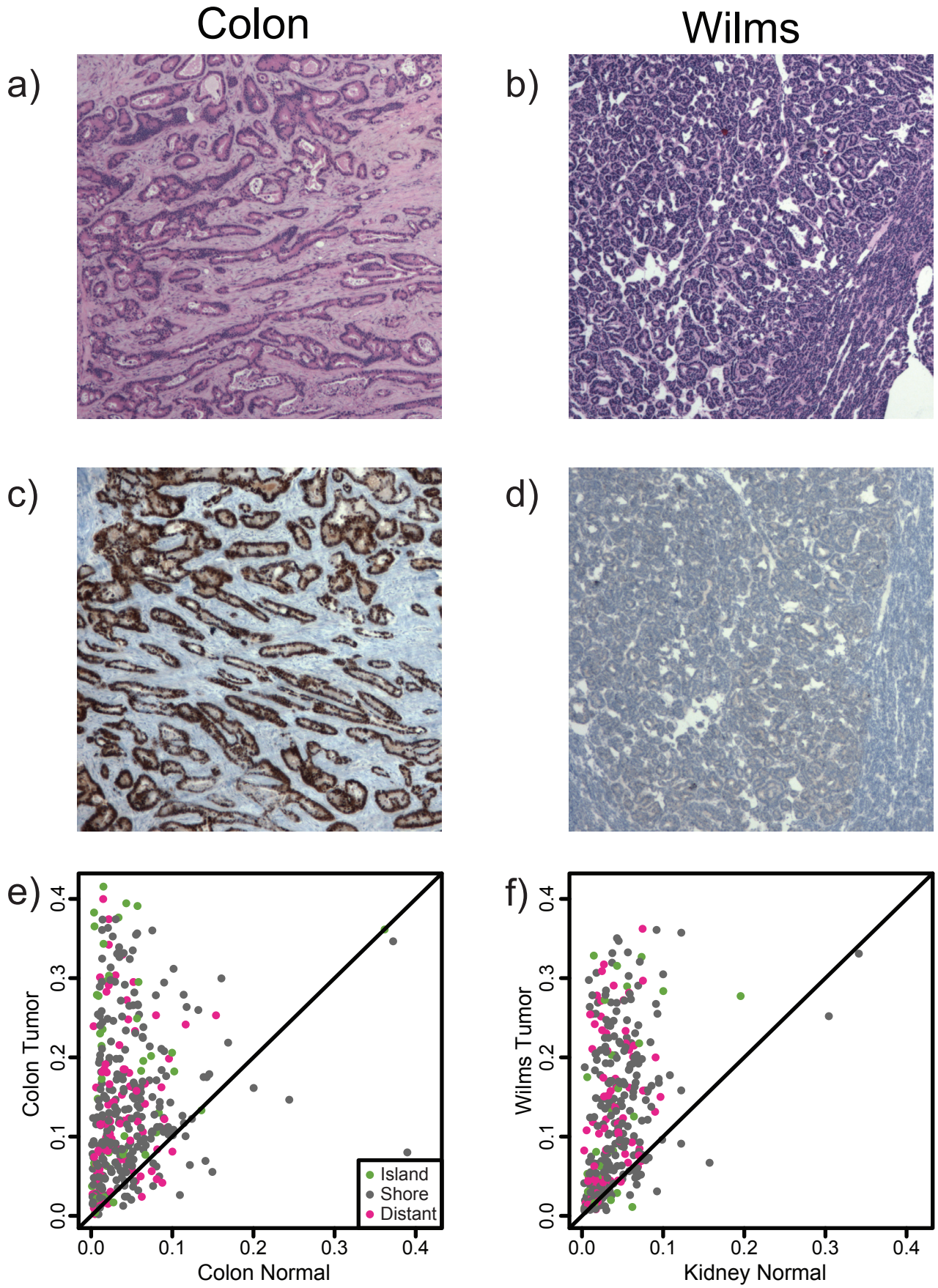




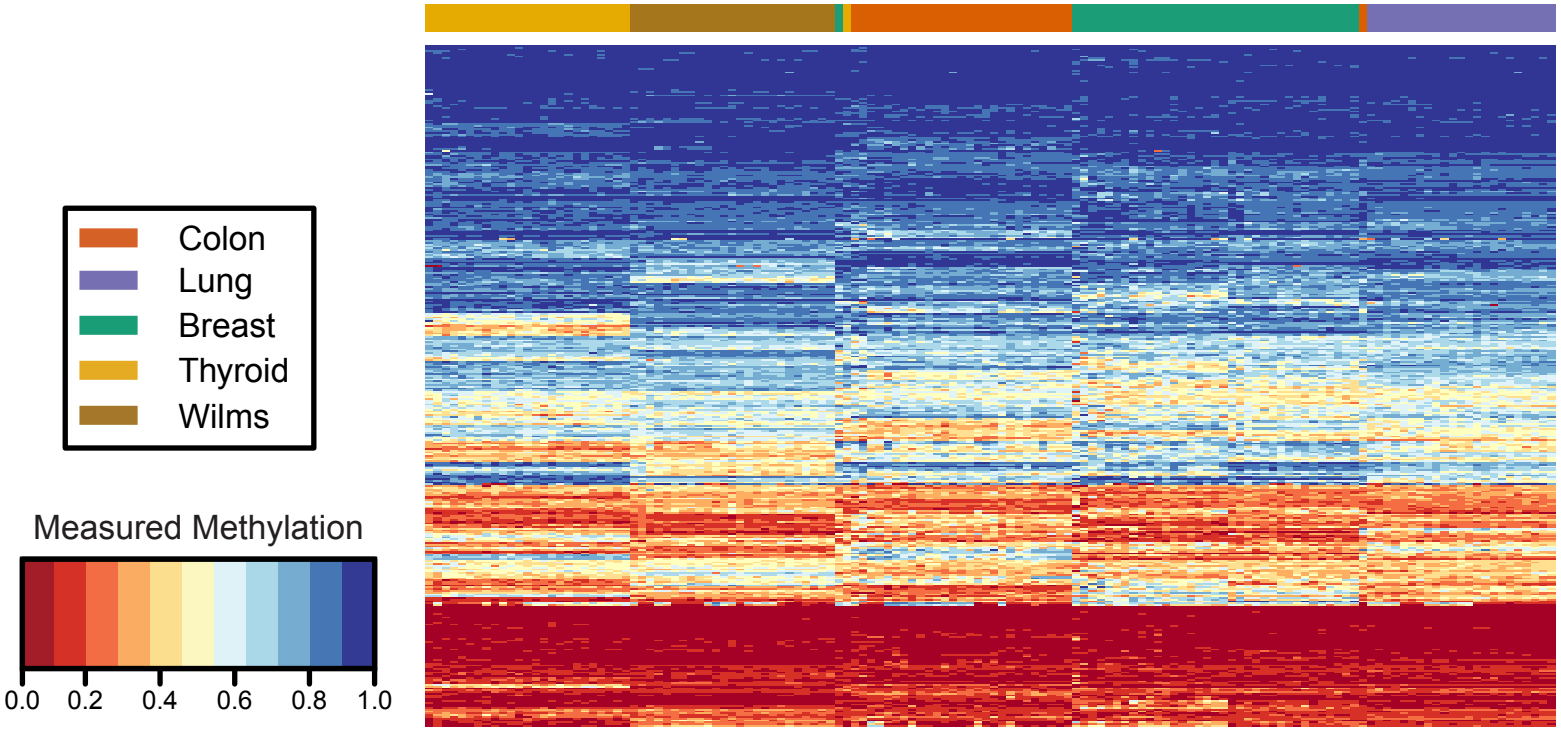
Supplementary Figure 4: Validation that increased variation in cancer is not caused by differences in age. Because 111 out of 122 samples were matched normal/tumor pairs, age was nearly balanced. However, because for lung, breast, and thyroid we included some samples that were not matched, we re-ran the analysis used for Figure 1 in the main manuscript after applying a regression model that corrected for age. **(a)** Lung **(b)** Breast **(c)** Thyroid - The conclusions are the same as described in the legend to Figure 1 in the main manuscript. **(d)** We also see the same results with this new analysis as in Figure 1f.



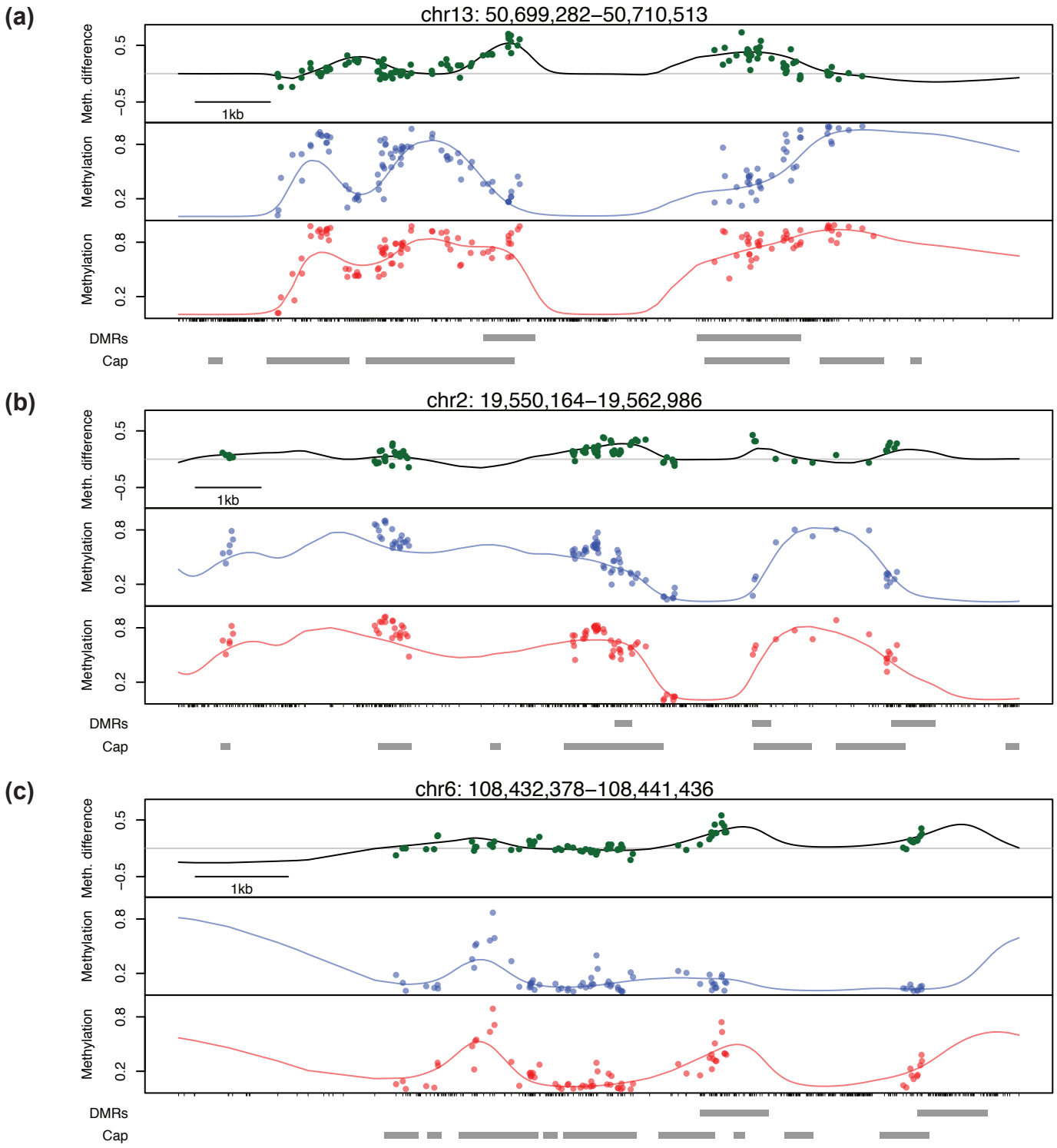
Supplementary Figure 5: Copy number variation versus methylation variance. Copy number variation of 5 colon tumor samples **a)** and 5 Wilms tumor samples **b)**. Chromosomes are colored in alternating green and black for clarity. Note that the Wilms samples generally show normal copy number, while the colon cancer samples demonstrate aneuploidy. Methylation levels measured at 384 CpG sites using the custom Illumina array exhibit an increase in across-sample variability in **c)** 5 colon normal/tumor matched samples and **d)** 5 kidney/Wilms tumor samples. Each panel shows the across-sample standard deviation of methylation level for each CpG in normal and matched cancer samples. The solid line is the identity line; CpGs above this line have greater variability in cancer. In both p53 postive and negative cancers, the vast majority of CpGs are above the solid line, indicating that the trend is independent of aneuploidy . Colors indicate the location of each CpG with respect to canonical annotated CpG islands.



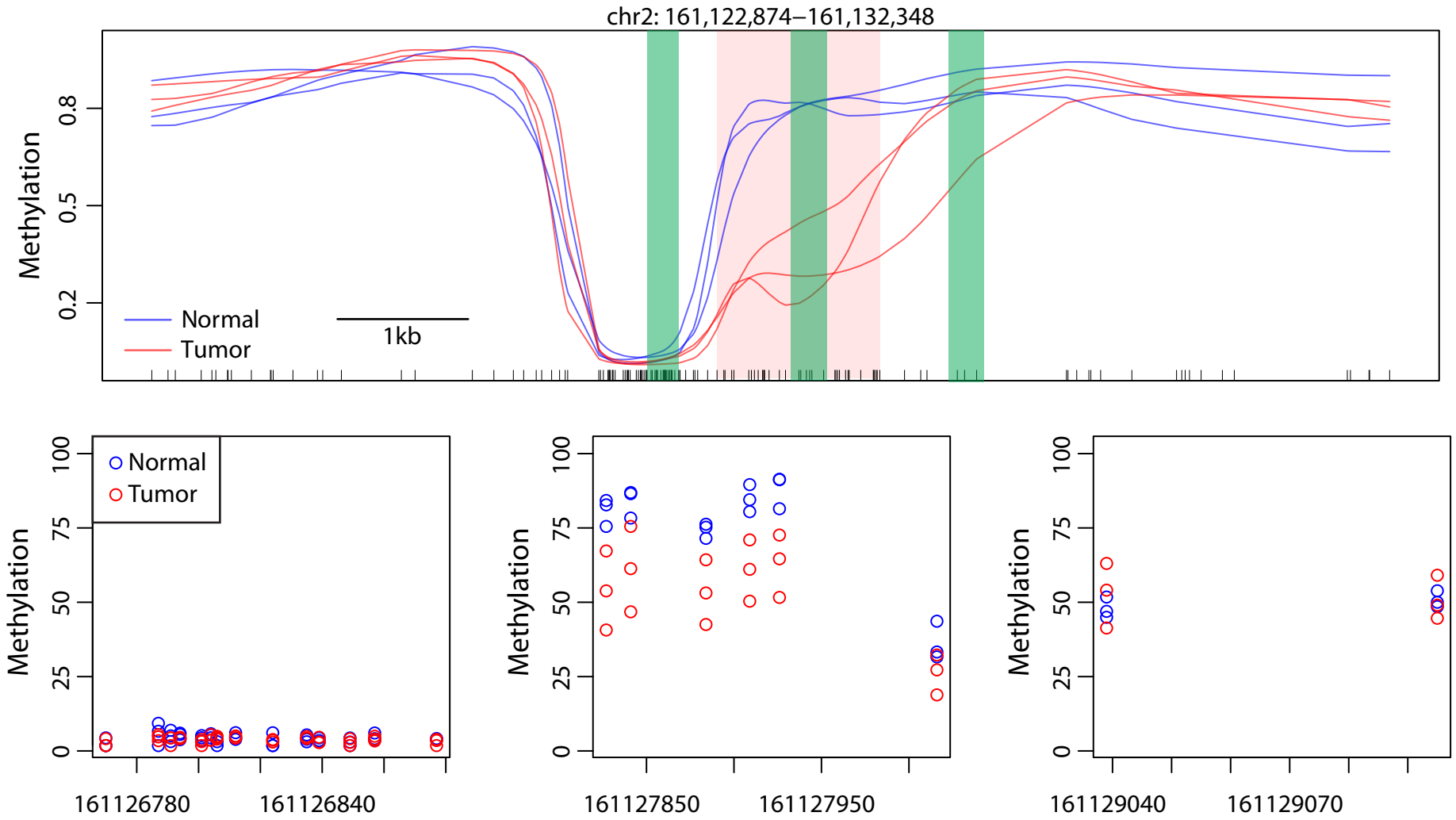
Supplementary Figure 6: Aberrant p53 expression versus methylation variance. H&E stains of **a)** Colon and **b)** Wilms tumor samples, all images taken at 4X magnification. p53 IHC stains for **c)** Colon and **d)** Wilms tumor samples; colon samples show a positive p53 stain, whereas Wilms samples do not. Methylation levels measured at 384 CpG sites using the custom Illumina array exhibit an increase in across-sample variability in **e)** 7 colon normal/tumor matched samples, all positive for p53 and **f)** 7 kidney/Wilms tumor samples, all negative for p53. Each panel shows the across-sample standard deviation of methylation level for each CpG in normal and matched cancer samples. The solid line is the identity line; CpGs above this line have greater variability in cancer. In both p53 postive and negative cancers, the vast majority of CpGs are above the solid line, indicating that the trend is independent of p53 status. Colors indicate the location of each CpG with respect to canonical annotated CpG islands.



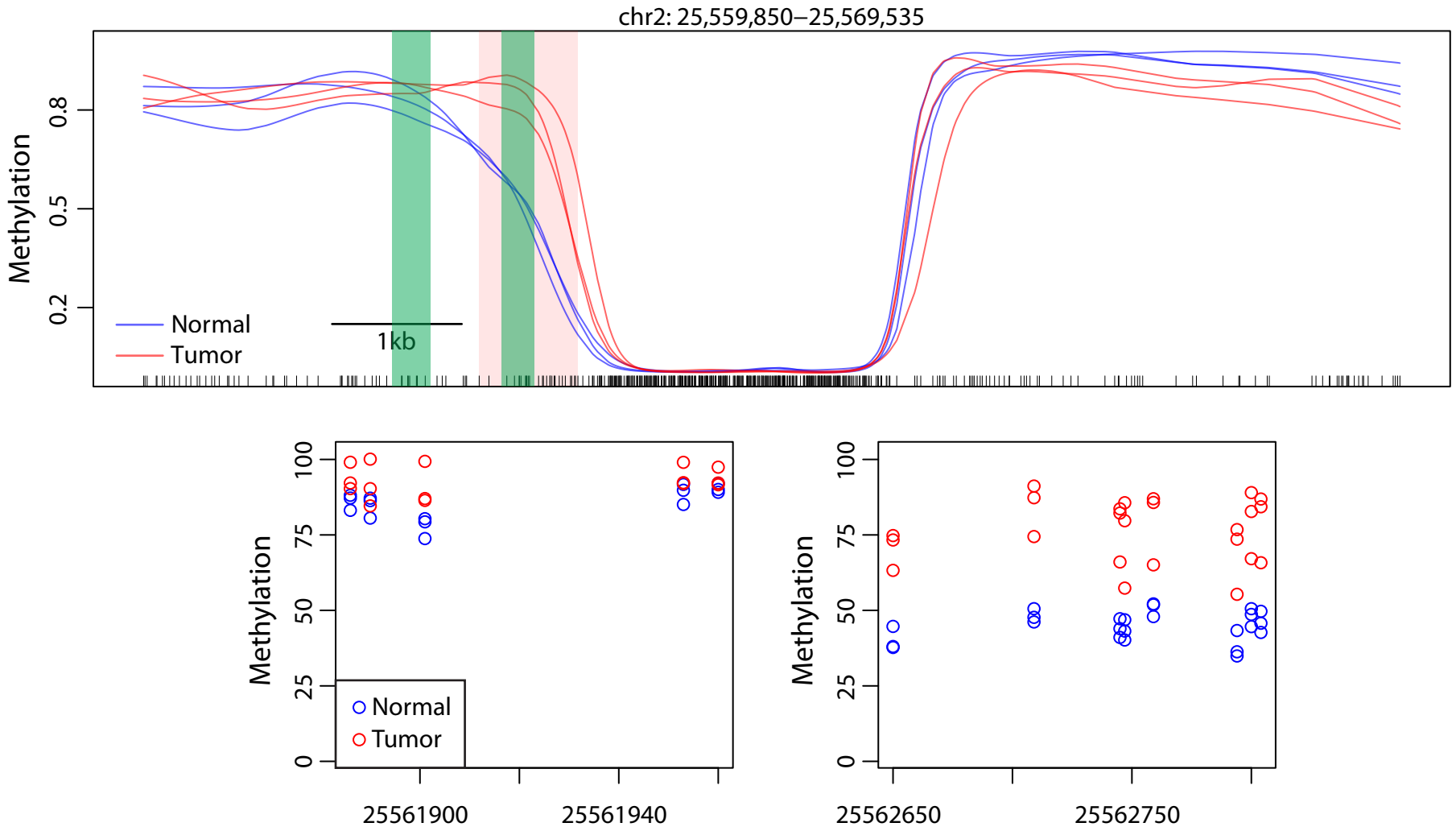
Supplementary Figure 7: Hierarchical cluster analysis of the normal samples using all probes. The heatmap of the methylation values for these clearly distinguishes the tissue types.



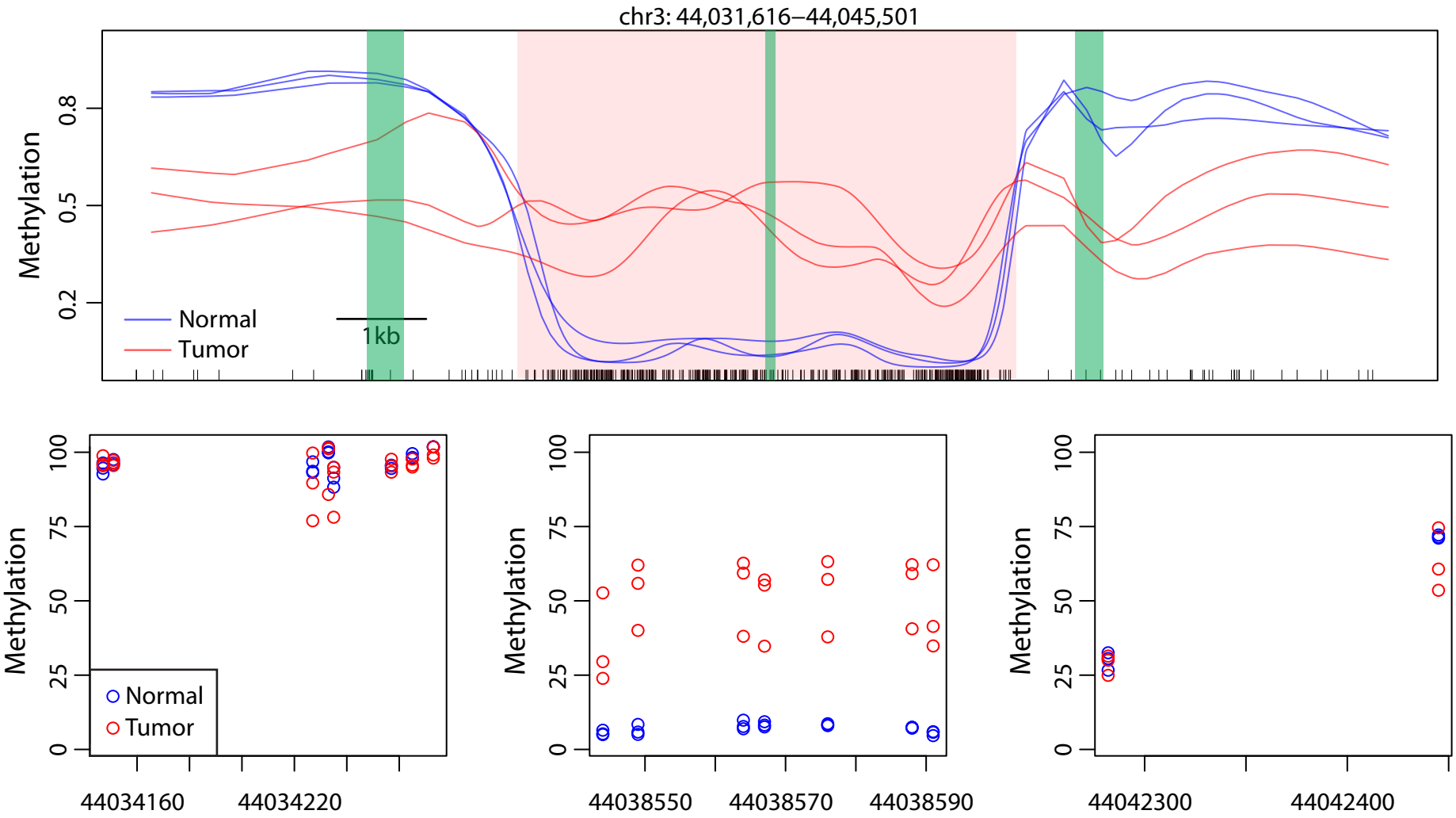
Supplementary Figure 8: Three examples illustrating the concordance between smoothed whole genome bisulfite sequencing and high-coverage (>30x) capture bisulfite sequencing. These three are representative of the genomic regions we used to corroborate our findings. **(a)** chr13: 50,699,282-50,710,513. The top panel shows the average methylation difference between the 3 cancer samples and the 3 normal samples. The black curve is the smoothed whole genome bisulfite data and the points are the single base resolution capture bisulfite data. The second panel shows the average of the methylation estimates for the 3 normal samples. The blue curve is the smoothed whole genome bisulfte data and the points are the single base resolution capture bisulfite data. The third panel shows the average of the methylation estimates for the 3 cancer samples. The red curve is the smoothed whole genome bisulfte data and the points are the single base resolution capture bisulfite data. The grey bars at the bottom show the locations of small DMRs and capture regions. **(b)** as **(a)** but for region chr2: 19,550,164-19,562,986. **(c)** as **(a)** but for region chr6: 108,432,378-108,441,436. Note that the curves go through the points, demonstrating the remarkable agreement between the smoothed whole-genome bisulfite data and the high-coverage capture bisulfite data. Additional regions are available at http://rafalab.jhsph.edu/cancer_seq/capture.pdf.



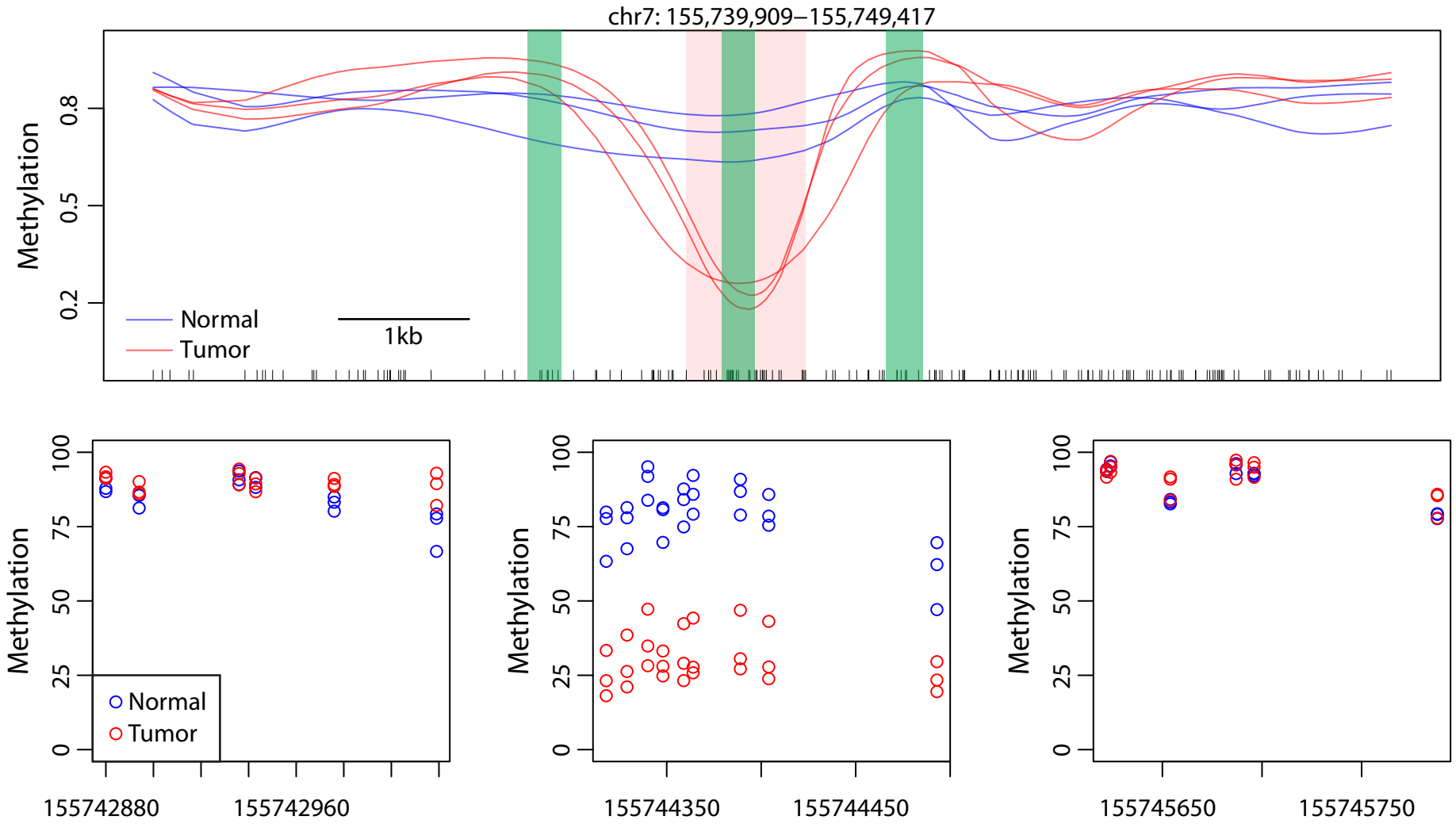
Supplementary Figure 9a: Bisulfite pyrosequencing of small DMRs. At top is pictured the same plot from Fig. 3a, showing the smoothed methylation values plotted against genomic location for the region. Cancer samples are plotted as red lines, and normal samples as blue lines with the DMR highlighted in pink. Regions which were bisulfite pyrosequenced are highlighted in green. Below, the methylation values from the three bisulfite pyrosequenced regions are plotted versus genomic position. Red circles are cancer samples, and blue circles are normal.



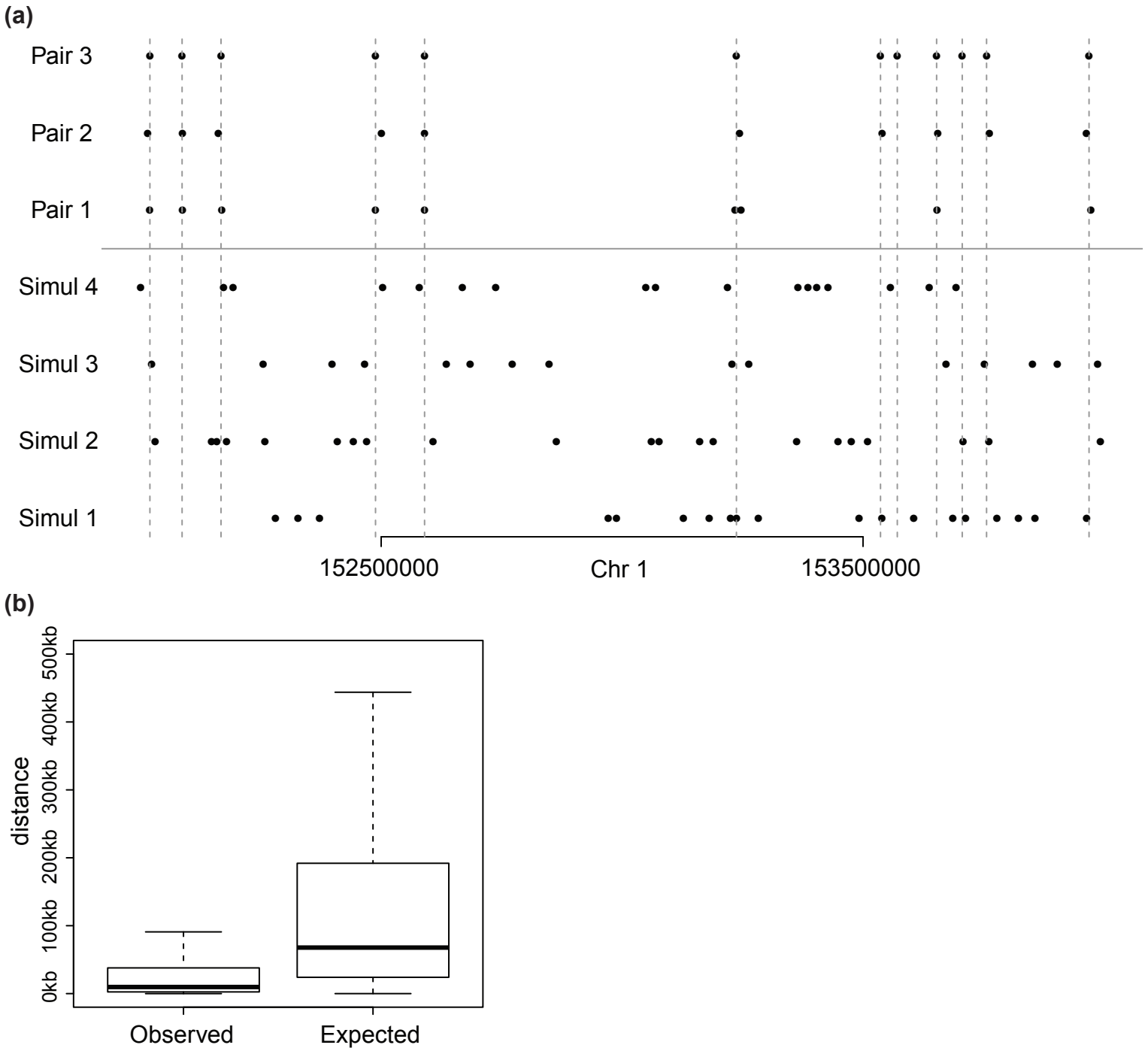
Supplementary Figure 9b: Bisulfite pyrosequencing of small DMRs. At top is pictured the same plot from Fig. 3b, showing the smoothed methylation values plotted against genomic location for the region. Cancer samples are plotted as red lines, and normal samples as blue lines with the DMR highlighted in pink. Regions which were bisulfite pyrosequenced are highlighted in green. Below, the methylation values from the three bisulfite pyrosequenced regions are plotted versus genomic position. Red circles are cancer samples, and blue circles are normal.



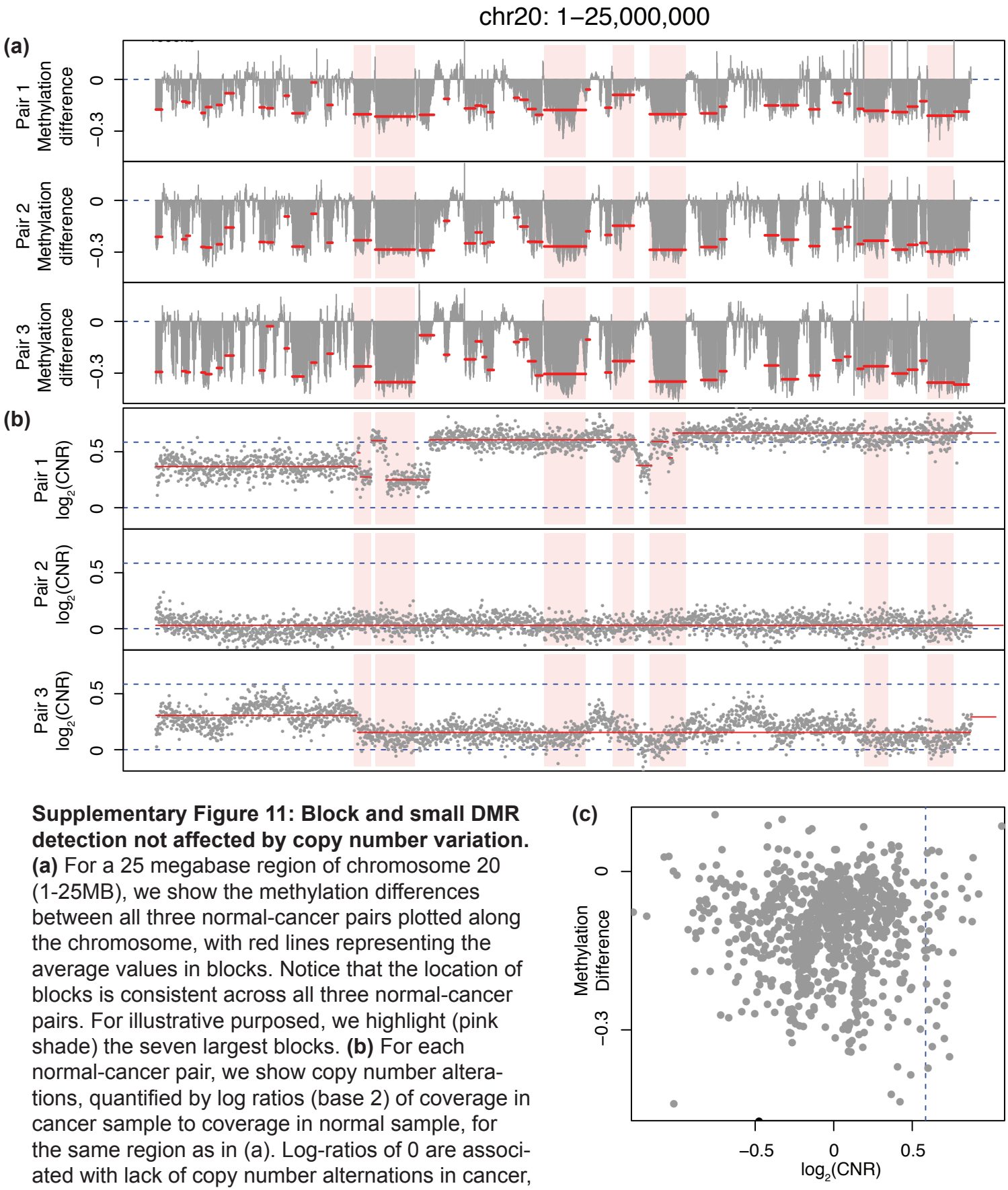
Supplementary Figure 9c: Bisulfite pyrosequencing of small DMRs. At top is pictured the same plot from Fig. 3c, showing the smoothed methylation values plotted against genomic location for the region. Cancer samples are plotted as red lines, and normal samples as blue lines with the DMR highlighted in pink. Regions which were bisulfite pyrosequenced are highlighted in green. Below, the methylation values from the three bisulfite pyrosequenced regions are plotted versus genomic position. Red circles are cancer samples, and blue circles are normal.



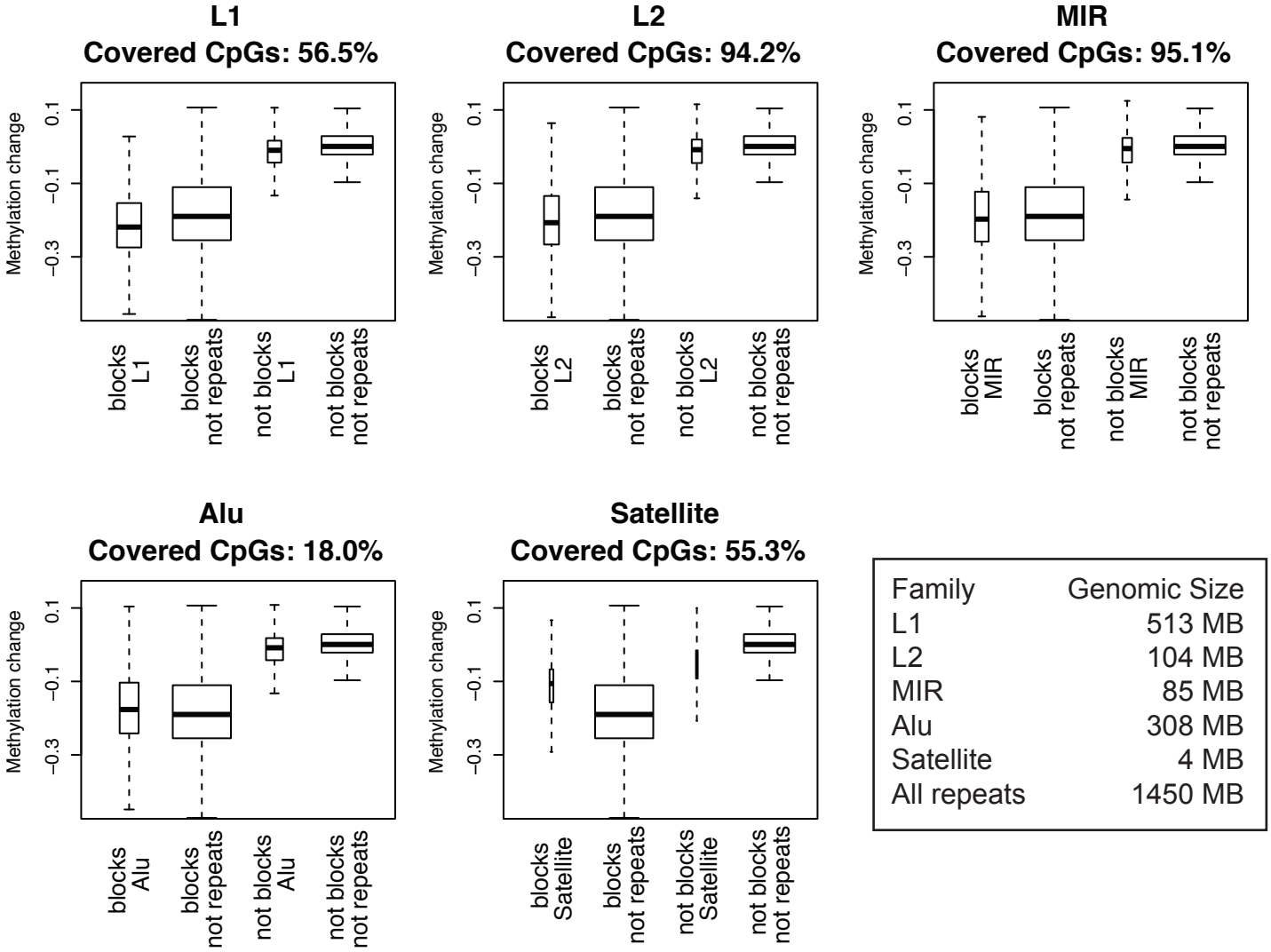
Supplementary Figure 9d: Bisulfite pyrosequencing of small DMRs. At top is pictured the same plot from Fig. 3d, showing the smoothed methylation values plotted against genomic location for the region. Cancer samples are plotted as red lines, and normal samples as blue lines with the DMR highlighted in pink. Regions which were bisulfite pyrosequenced are highlighted in green. Below, the methylation values from the three bisulfite pyrosequenced regions are plotted versus genomic position. Red circles are cancer samples, and blue circles are normal.



Supplementary Figure 10: Simulations show that block locations co-occur. (a) Start locations of blocks for each of the 3 cancer-normal pairs are shown for a 2MB region on chromosome 1. Also shown are 4 simulated sets of block start positions. Vertical lines represent the start locations of blocks for cancer-normal pair 3. **(b)** For each of the 3 cancer-normal pairs we computed the distance from the observed start position of each sample-specific block to the closest start position in the other pairs. The boxplot on the left shows the distribution of these distances, pooled across all possible comparisons. The boxplot on the right shows the expected distribution of distances under the null hypothesis that the block start positions do not agree. The smaller values seen in the left boxplot demonstrates that the start positions of the sample-specific blocks co-occur much more frequently than expected by chance.

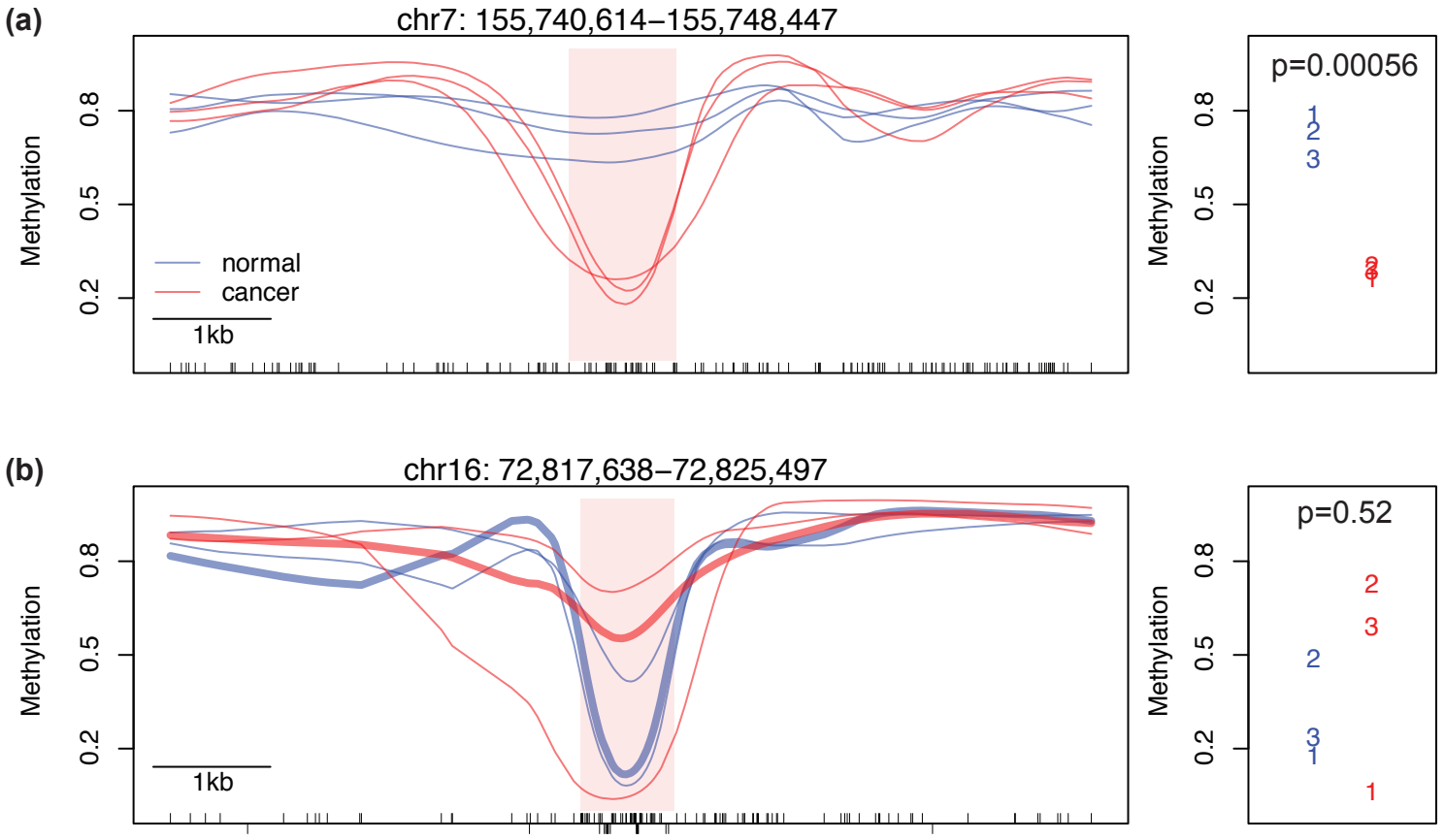


Supplementary Figure 11: Block and small DMR detection not affected by copy number variation. **(a)** For a 25 megabase region of chromosome 20 (1-25MB), we show the methylation differences between all three normal-cancer pairs plotted along the chromosome, with red lines representing the average values in blocks. Notice that the location of blocks is consistent across all three normal-cancer pairs. For illustrative purposed, we highlight (pink shade) the seven largest blocks. **(b)** For each normal-cancer pair, we show copy number alterations, quantified by log ratios (base 2) of coverage in cancer sample to coverage in normal sample, for the same region as in (a). Log-ratios of 0 are associated with lack of copy number alternations in cancer, while values larger or equal to $\log_2(3/2)$ (dashed line) are associated with gain of copy number in cancer. The red lines represent segments obtained with the CBS algorithm (described in the Supplementary Note). Notice that each sample shows different copy number alterations. **(c)** Differences in methylation are plotted against differences in copy number (log-ratios) for the entire genome. Specifically, for each of the segments detected by CBS, we computed the average difference in methylation and the average log-ratio associated with copy number alteration. We did this for each sample and combined all the points in one scatter-plot.

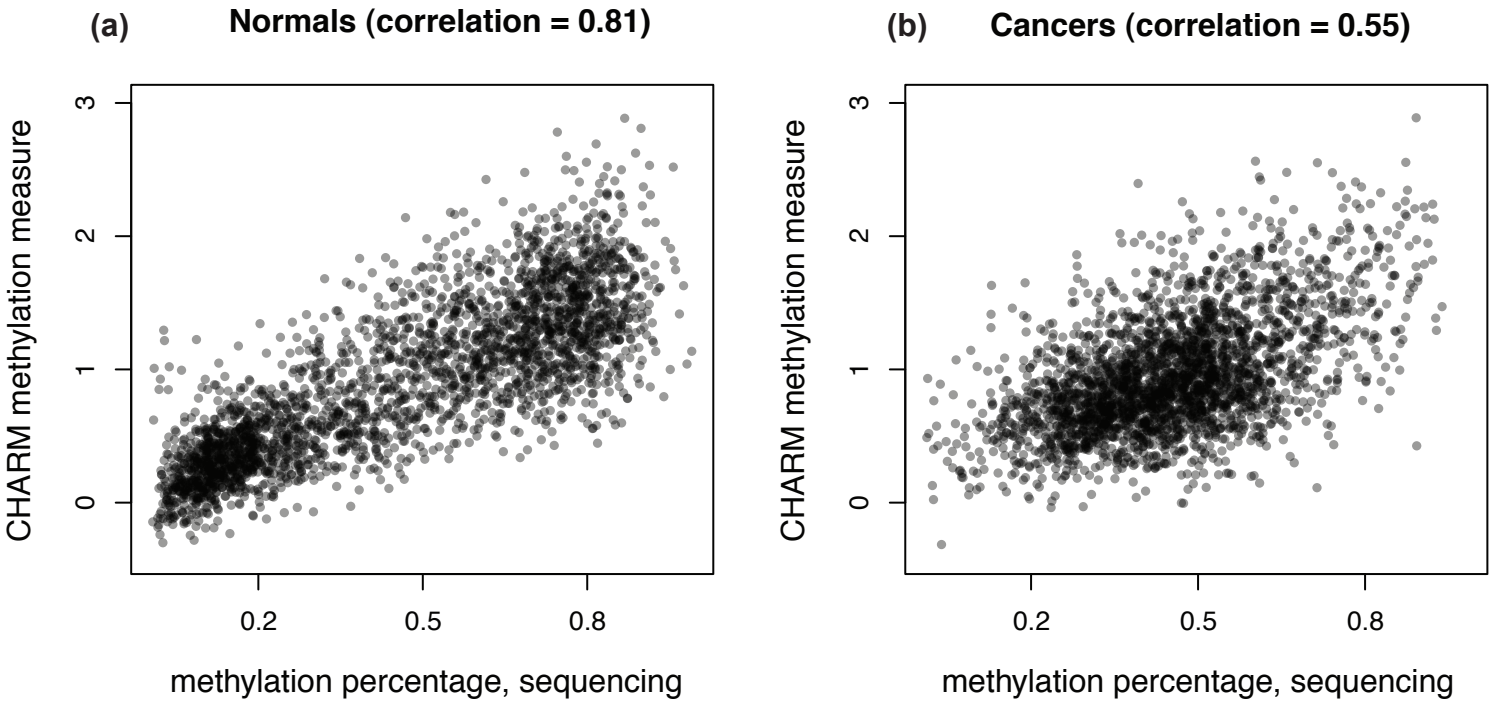


Family	Genomic Size
L1	513 MB
L2	104 MB
MIR	85 MB
Alu	308 MB
Satellite	4 MB
All repeats	1450 MB

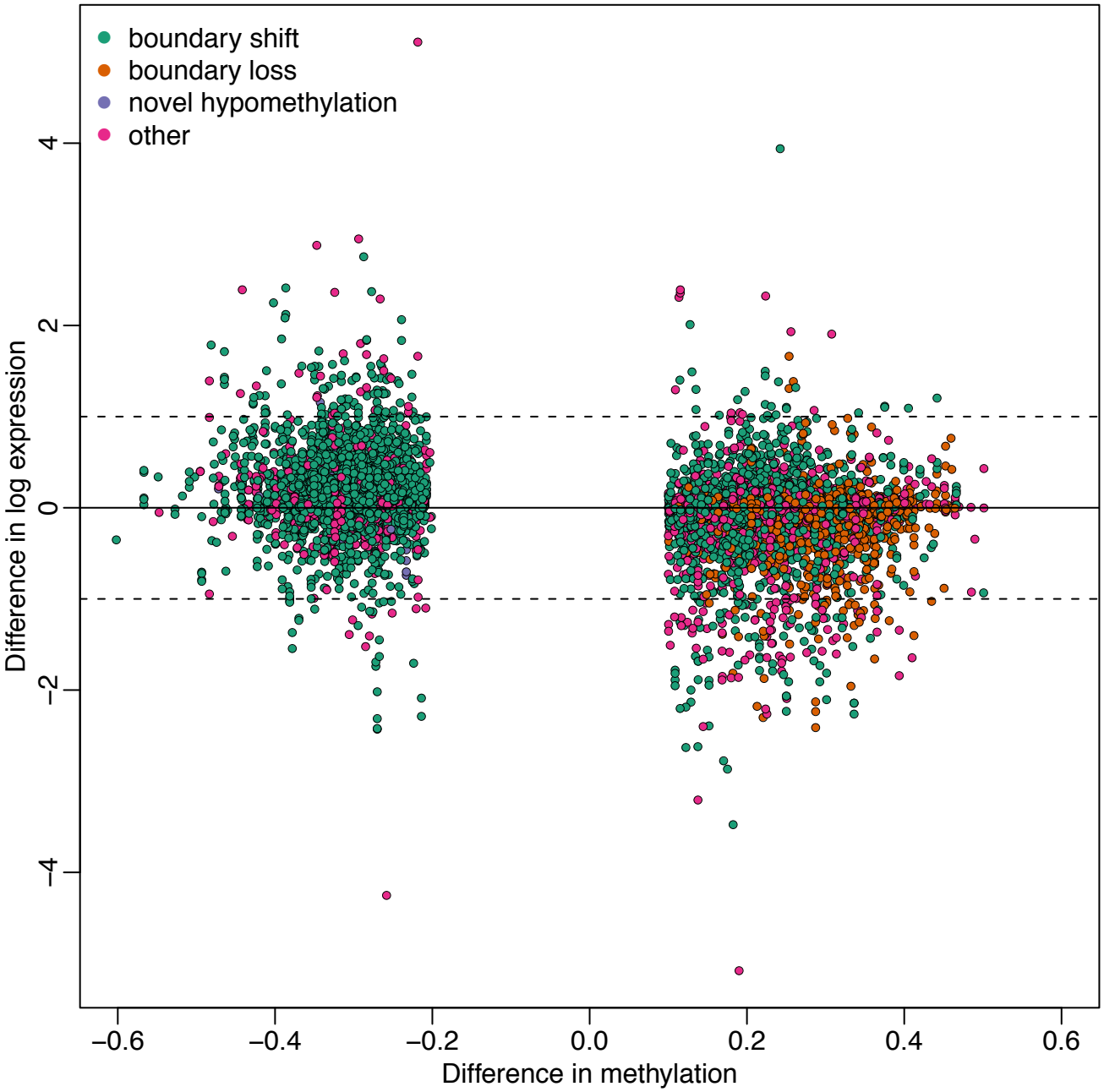
Supplementary Figure 12: Methylation changes of selected repetitive DNA families. Distribution of methylation differences between cancer and normal samples stratified by repeat family and inclusion in blocks. CpGs outside all repetive elements were used as controls. Above each graph is the percentage of covered CpGs in each repeat family.



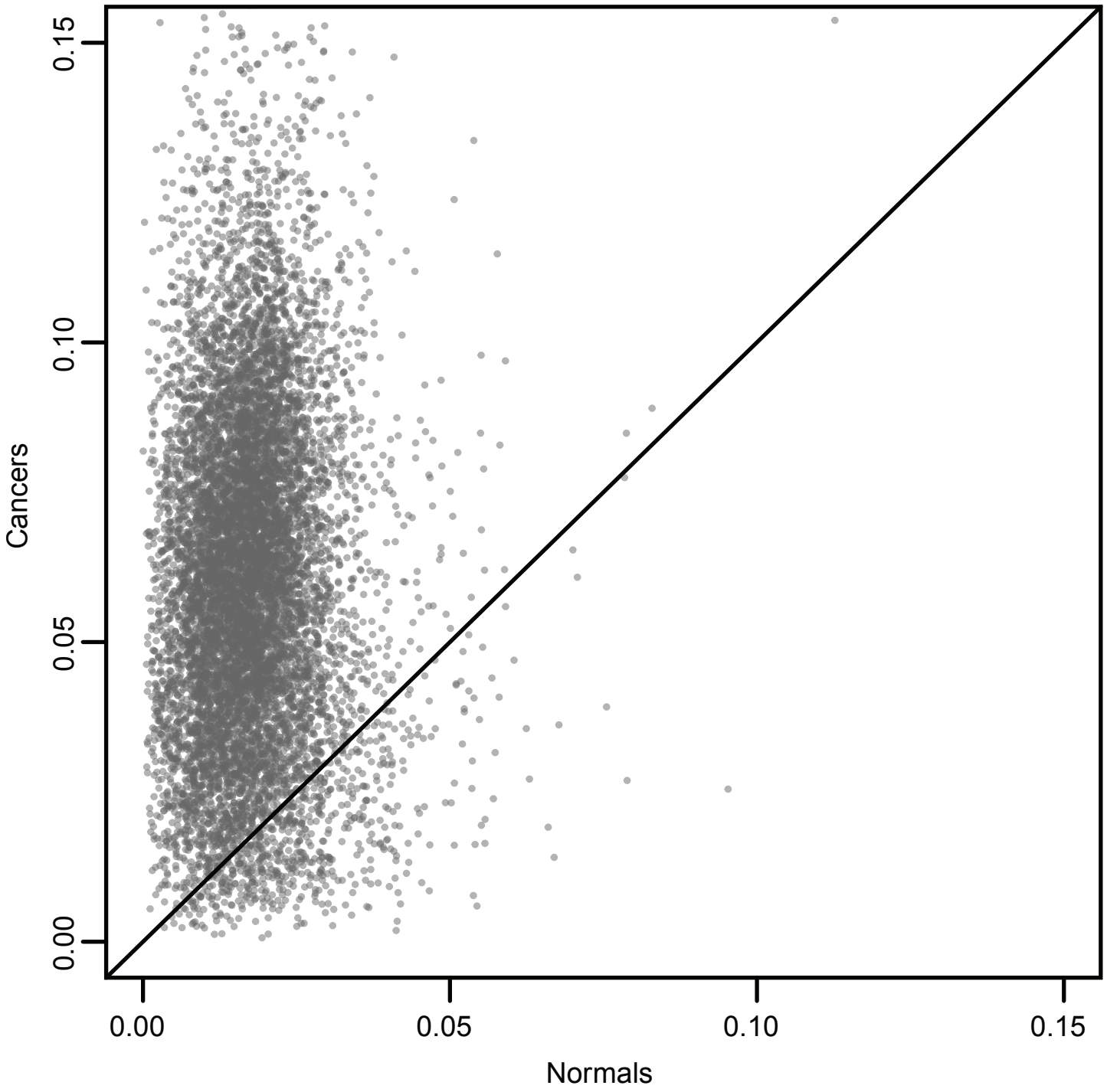
Supplementary Figure 13: The need for biological replication for detecting DMRs. (a) In the left panel, we show methylation patterns for three normal samples (blue) and matched cancers (red). The detected DMR is shaded in pink. In the right panel, we show the average methylation values within the DMR for the three paired samples (normal in blue, cancer in red, the matched sample pairs indicated by numbers). We performed a t-test for the difference between normal and cancer and obtained a p value of 0.0056. **(b)** We show the same analysis as (a) for a region in which if we had only analyzed normal-cancer pair 3 (thick lines), there would appear to be a methylation difference between cancer and normal. However, the p value when all three samples are compared is 0.52. Notice that our methylation estimates are very precise (standard error for points in right panel are < 0.02). Therefore the differences we see are biological not technical, and will be seen regardless of the measurement technology.



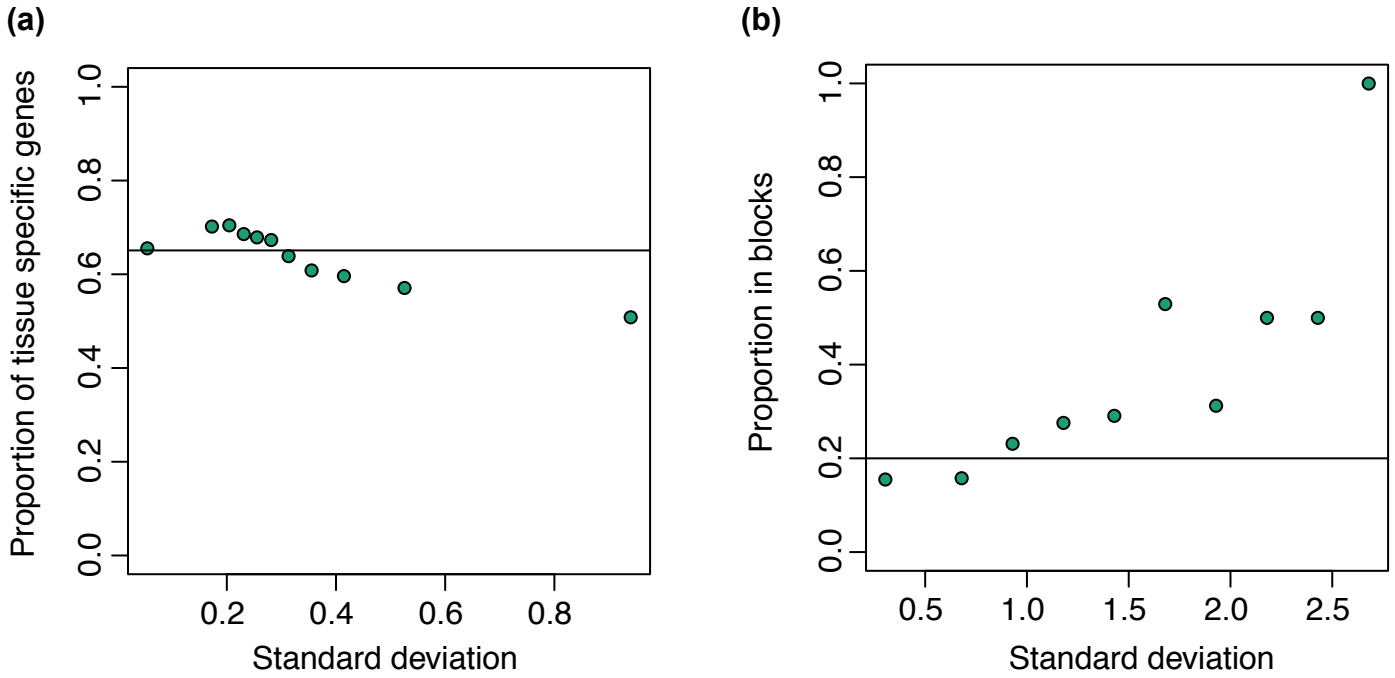
Supplementary Figure 14: Comparison of CHARM microarray data and bisulfite sequencing for measuring methylation. Average methylation level from previously published CHARM microarray data (Irizarry et al., 2009) (y-axis) is plotted versus the average methylation obtained from high-frequency smoothed bisulfite sequencing data. Each point represents one of the cDMR regions originally identified in (Irizarry et al., 2009). **(a)** Normal and **(b)** Cancer samples. Note the high degree of correlation between CHARM and sequencing.



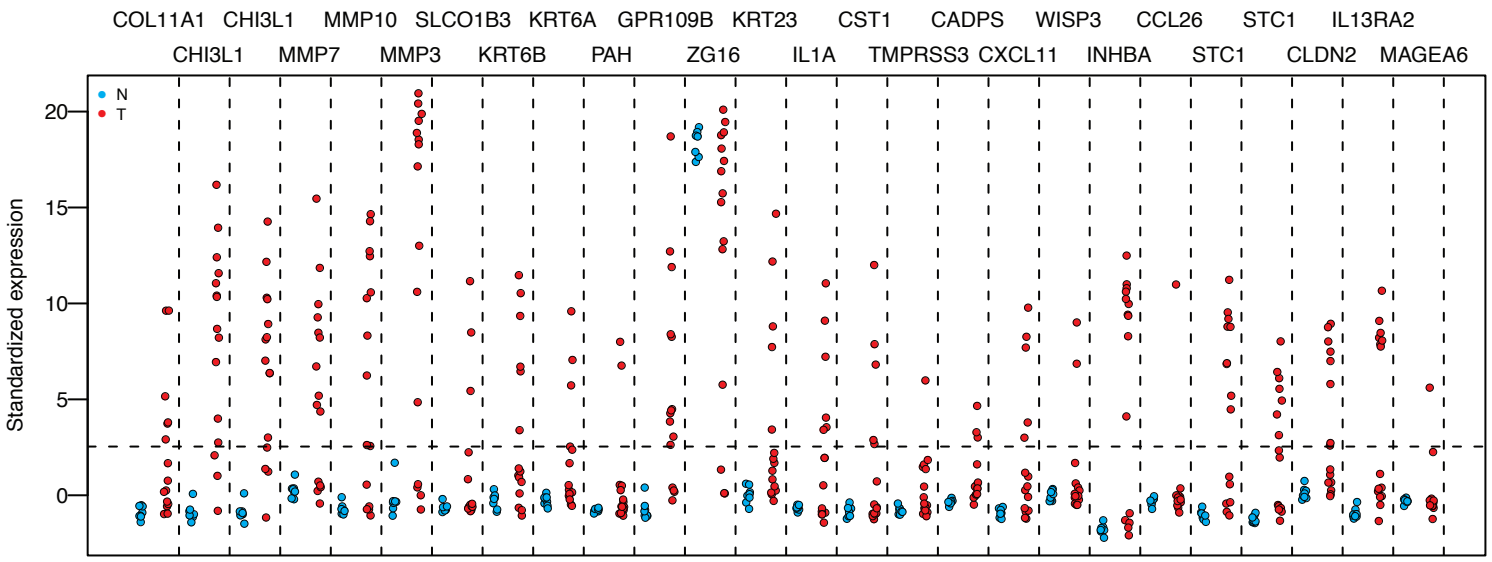
Supplementary Figure 15: Inverse correlation of gene expression with methylation at small DMRs. Average log gene expression values, obtained from GEO dataset GSE8671, plotted versus the average difference in methylation of a nearby small DMR. We considered a gene and a small DMR associated if the DMR was within 2,000 bp of the transcription start site of the gene; 6,869 genes mapped to a DMR in this way. Different types of small DMRs are indicated by color, boundary shift (green), boundary loss (orange), novel hypomethylation (purple) and other (pink). The dashed lines represent a fold change of 2 in the gene expression comparison.



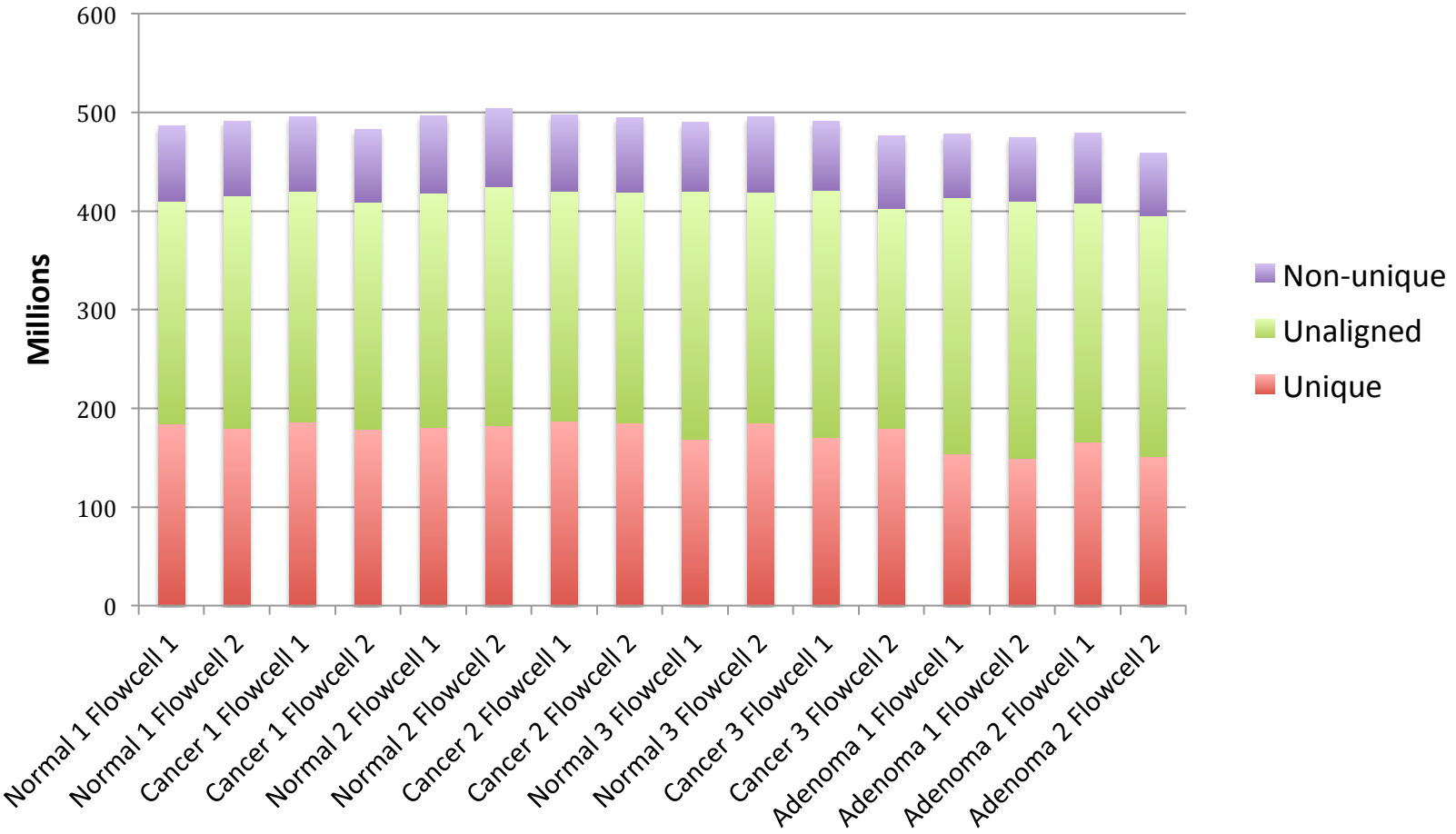
Supplemental Figure 16: Increased variation in methylation between normal and cancer samples in blocks. Across-sample standard deviation of methylation level for each block of normal versus cancer samples. Average methylation levels were computed for each block using high frequency smoothed SOLiD bisulfite sequencing data. The solid line is the identity line; CpGs above this line have greater variability in cancer. As in Figure 1, the vast majority of blocks show an increased variation in cancer compared to normal samples.



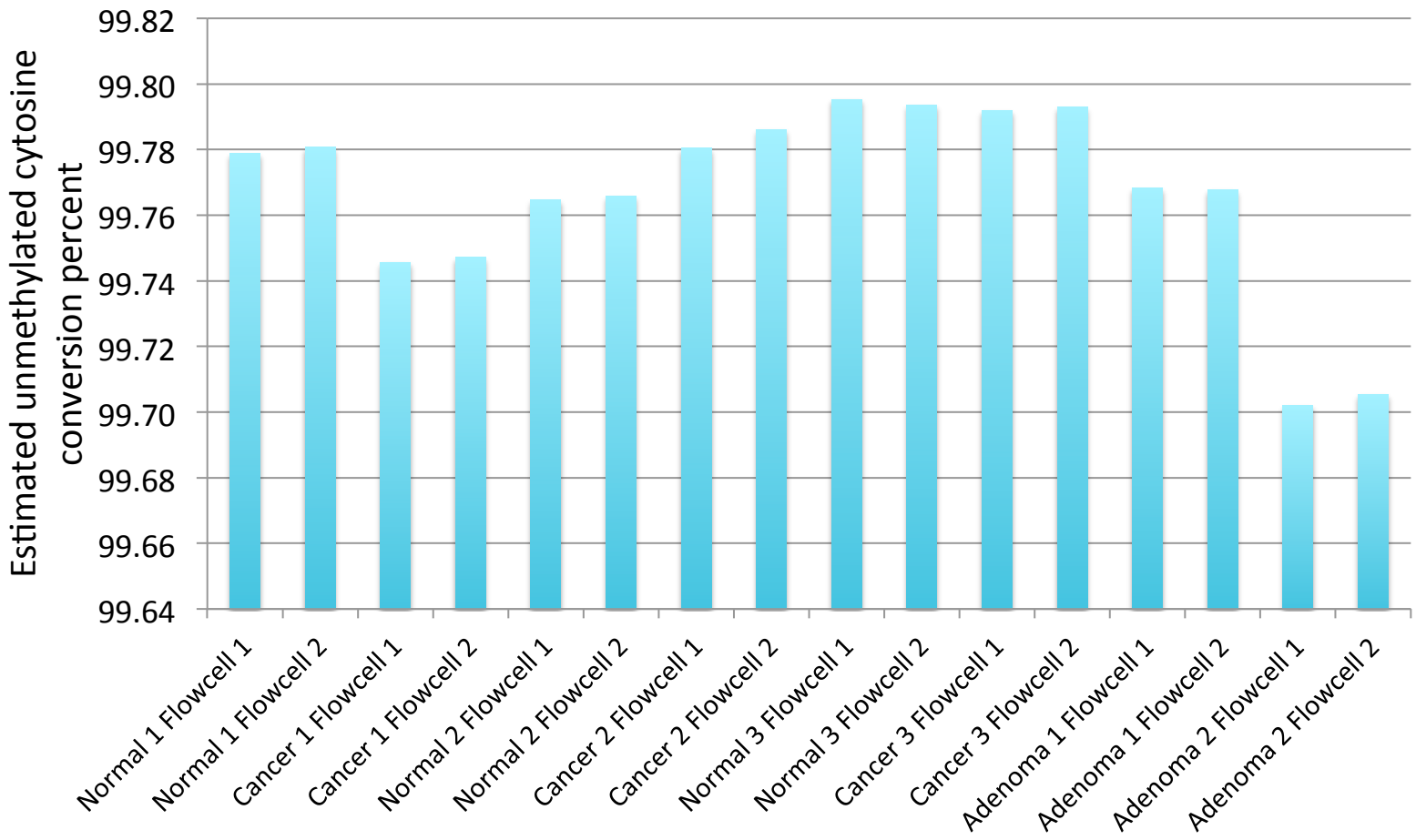
Supplementary Figure 17: Proportion of hyper-variable genes inside blocks. Genes represented in the Affymetrix HGU-133Aplus array were stratified by their across-sample standard deviation in cancer into 10 bins. Expression from a colon cancer study was used. **(a)** For each bin, the proportion of tissue-specific genes was computed for genes with their transcription start site (TSS) outside of blocks. Here we plot these proportions against the average standard deviation of the respective bin. The horizontal line represents the proportion of TSS represented in the microarray that are outside hypomethylated blocks and are tissue-specific genes. **(b)** For each bin, the proportion of genes with their transcription start site (TSS) in a hypomethylated blocks was calculated. Here we plot these proportions against the average standard deviation of the respective bin. The horizontal line represents the proportion of TSS represented in the microarray that are inside hypomethylated blocks.



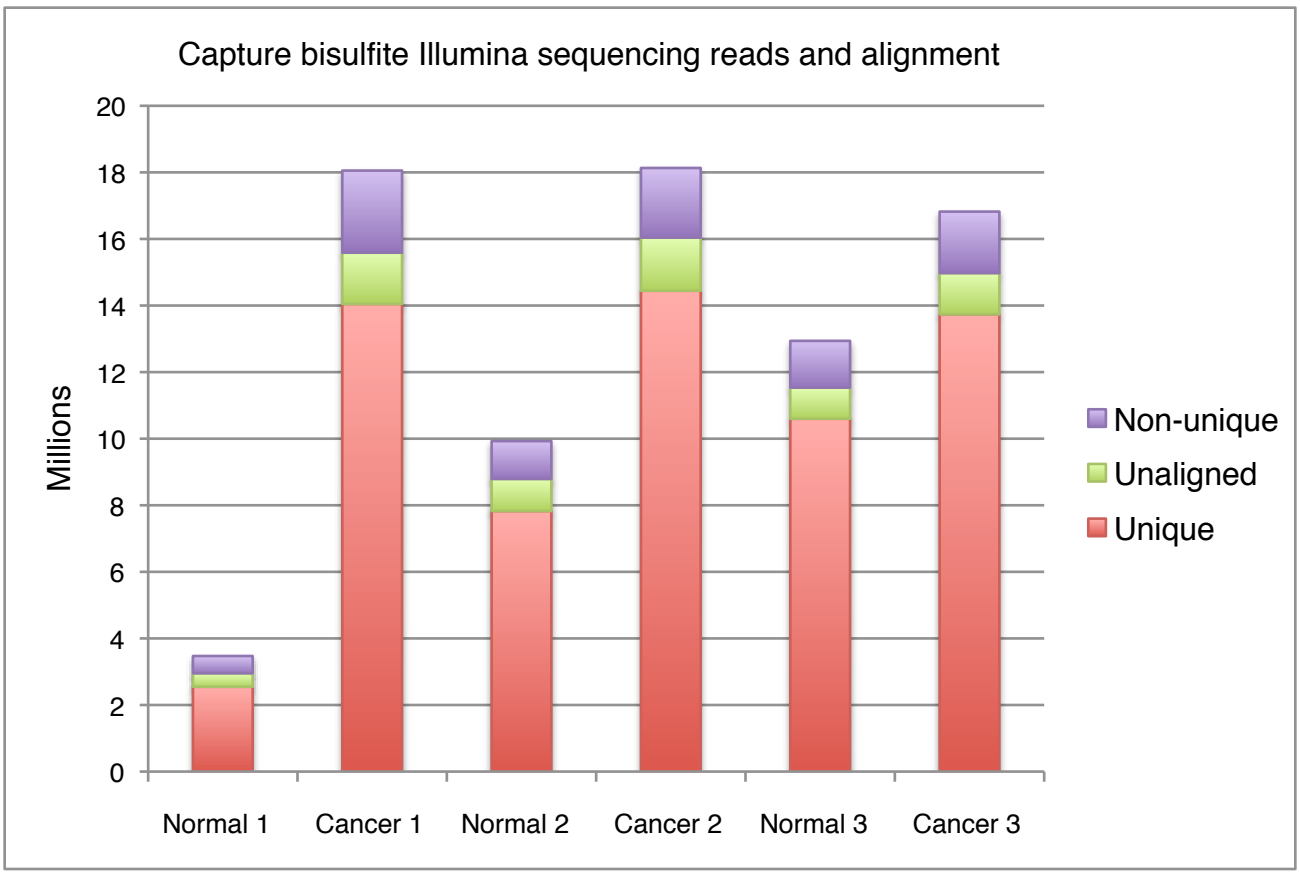
Supplementary Figure 18: Hypervariable gene expression in cancer in hypomethylated blocks. Standardized (using gene expression barcode) expression values for the 26 of the 50 most hypervariable genes in cancer which are within hypomethylated block regions. Genes with standardized expression values below 2.54 (dotted horizontal line) are determined to be silenced by the barcode method (Zilliox and Irizarry, 2007). Expression values for each gene, separated by vertical dotted lines, from dataset GSE4183 are plotted for normal (blue) and cancer (red) samples. Note there is consistent expression silencing in normal samples compared to variable expression in cancer samples. This plot is similar to Figure 5b, but for a different expression dataset.



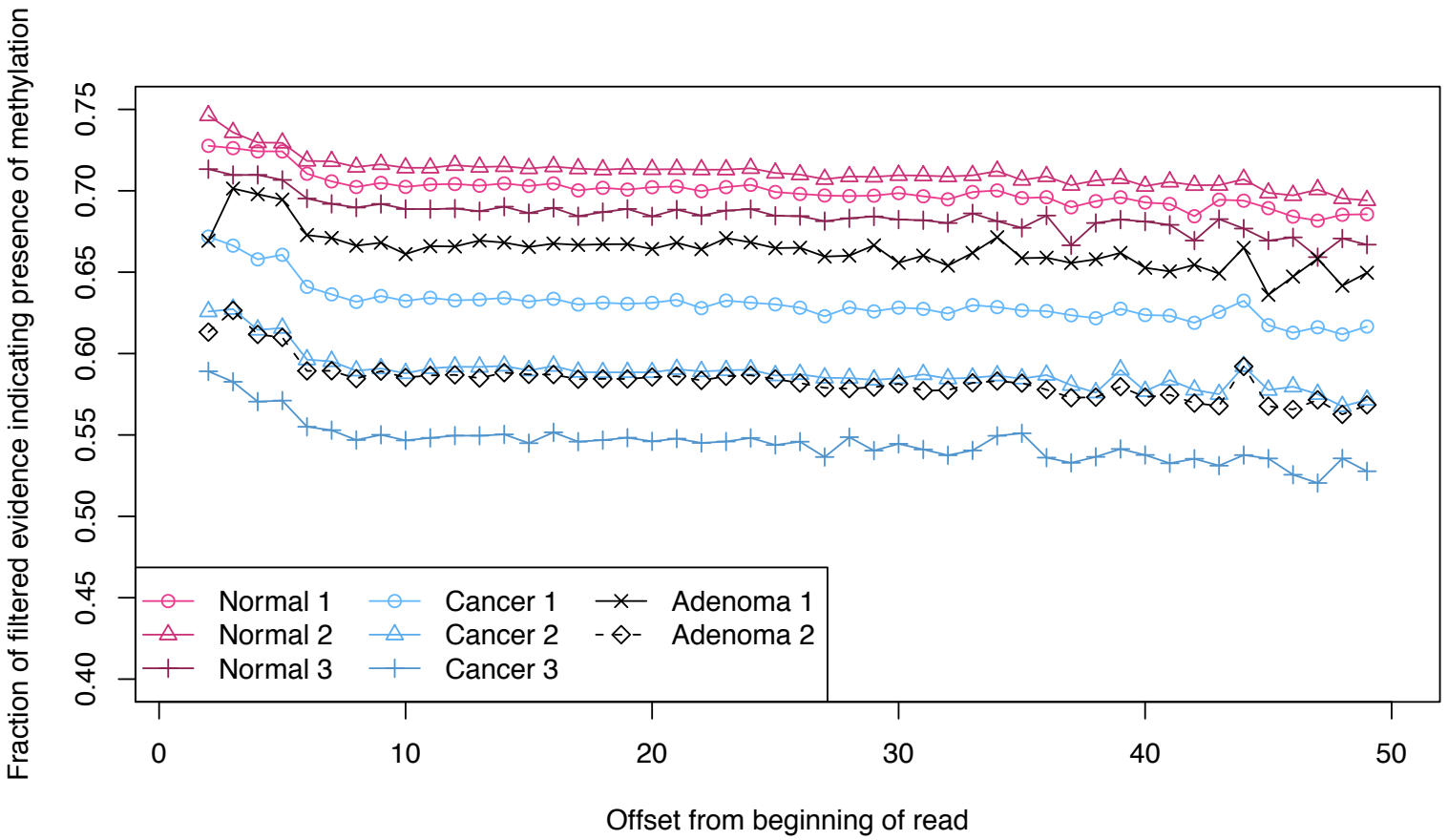
Supplementary Figure 19: Number of reads and alignments obtained from SOLiD 3+ bisulfite sequencing. The stacked bar chart illustrates the number of reads sequenced per flowcell, with the colors indicating uniquely reads (red), unaligned reads (green) and non-uniquely aligned reads (purple). A total of 7.79 billion reads were obtained from 8 runs (16 flowcells) of a SOLiD 3+ instrument.



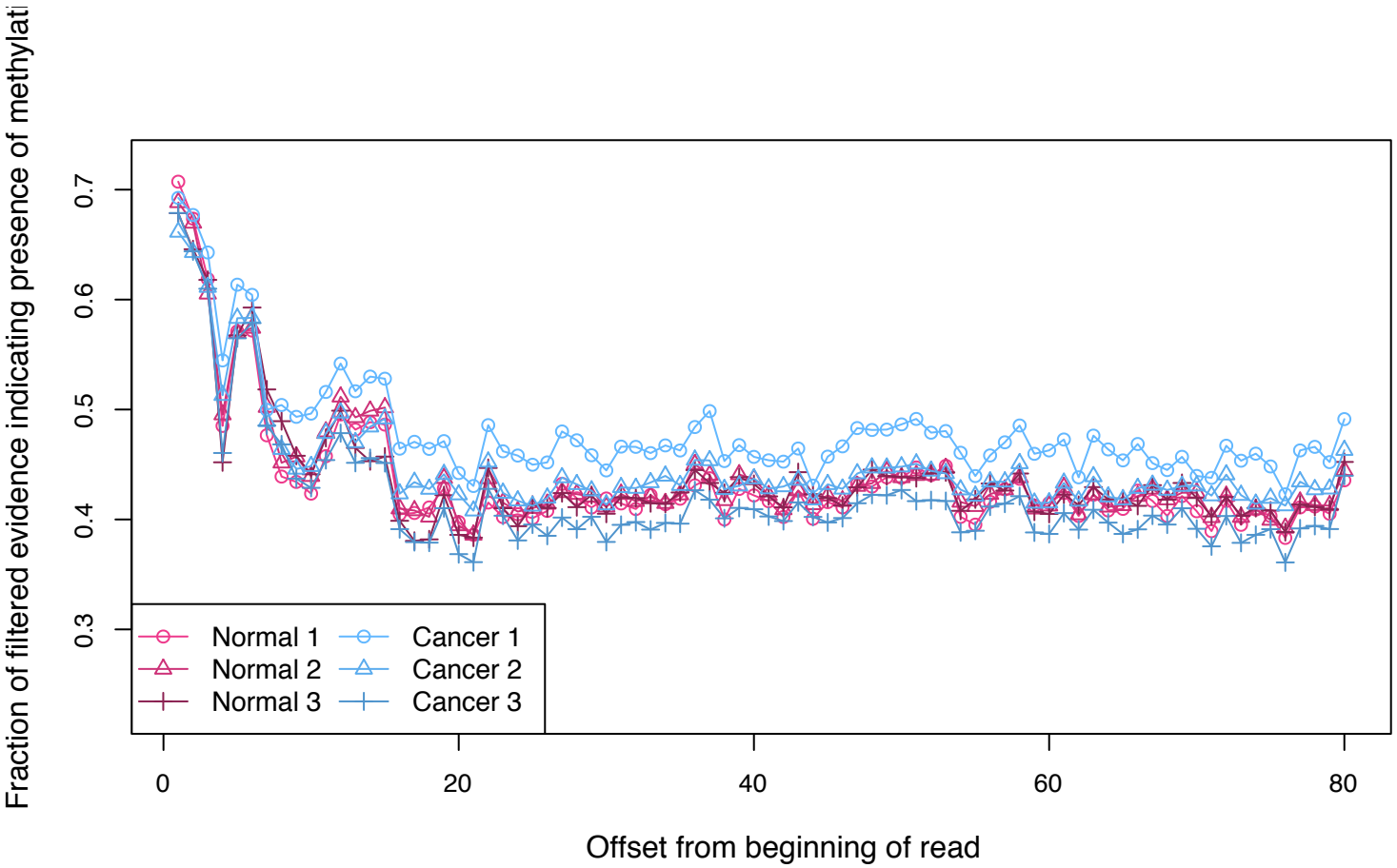
Supplementary Figure 20: Estimated unmethylated cytosine conversion rate per sample. Bisulfite conversion efficiency is plotted per SOLiD 3+ flowcell. Conversion efficiency is estimated as the fraction of high-quality evidence aligning to CpG cytosines in the unmethylated λ phage genome that indicates lack of methylation.



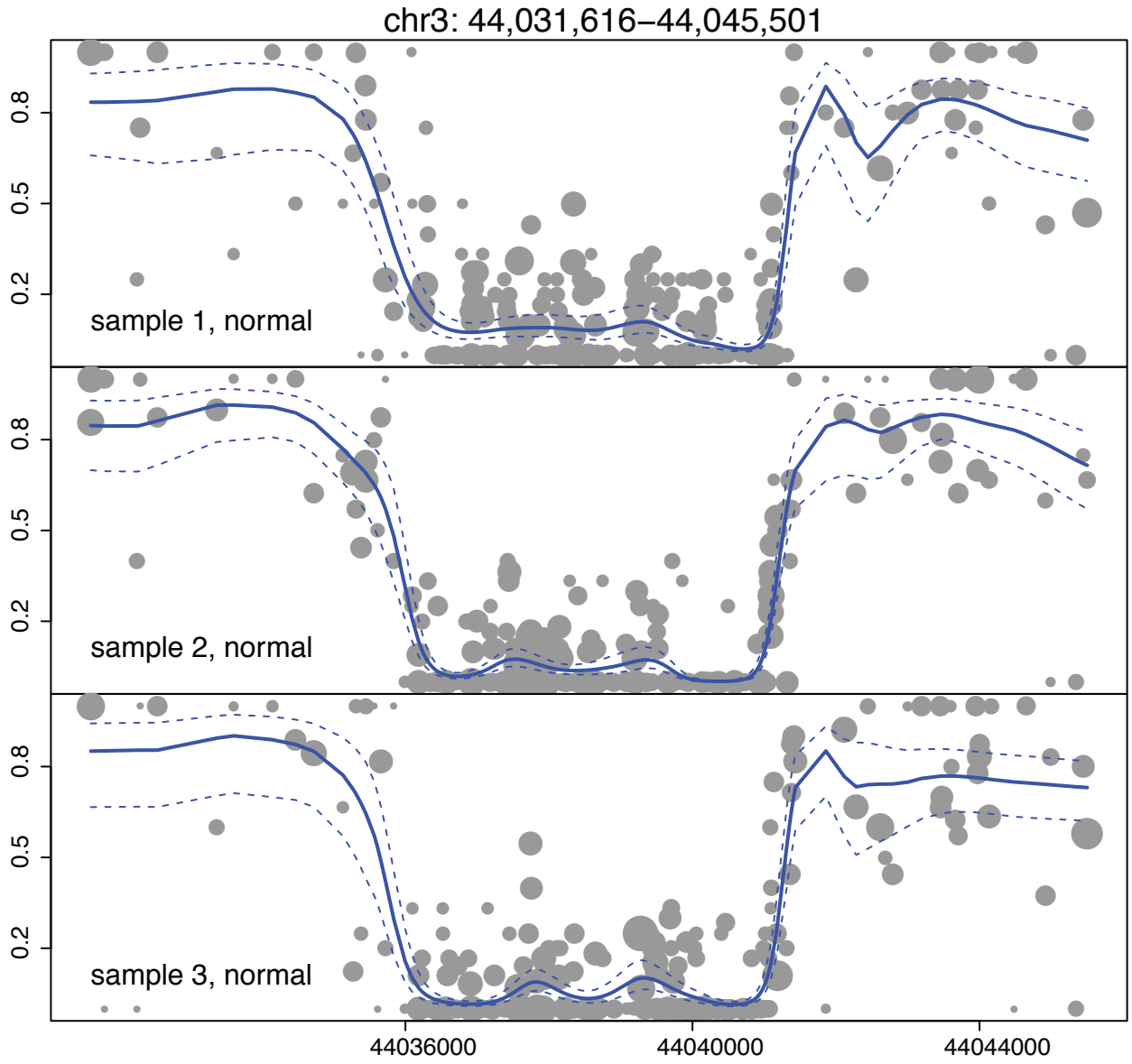
Supplementary Figure 21: Number of reads and alignments obtained from Illumina GA II bisulfite capture sequencing. The stacked bar chart illustrates the number of reads sequenced per sample, with the colors indicating uniquely reads (red), unaligned reads (green) and non-uniquely aligned reads (purple). A total of 79.3 million reads were obtained.



Supplementary Figure 22: SOLiD 3+ Read position bias in evidence for methylation. The horizontal axis represents an offset into the nucleotide alignment from the 5' end. The vertical axis represents the fraction of filtered CpG methylation evidence from that offset that indicates that methylation is present. Only reads aligning uniquely to the GRCh37 human genome assembly are considered. In a perfect assay, the fraction should be independent of alignment offset and each line should be flat and horizontal. In practice, the lines are not flat due to sequencing error and other noise arising from sample preparation and alignment.



Supplementary Figure 23: Read position bias in evidence for methylation in capture bisulfite data sequenced on the Illumina GA II instrument. The horizontal axis represents an offset into the nucleotide alignment from the 5' end. The vertical axis represents the fraction of filtered CpG methylation evidence from that offset that indicates that methylation is present. Only reads aligning uniquely to the GRCh37 human genome assembly are considered. In a perfect assay, the fraction should be independent of alignment offset and each line should be flat and horizontal. In practice, the lines are not flat due to sequencing error and other noise arising from sample preparation and alignment. Based on this plot the first 15 bases of the reads were trimmed before further analysis.



Supplementary Figure 24: Precise methylation estimates obtained by high-frequency smoothing. The circles represent the single CpG estimates of methylation, which are plotted against the CpG location. The areas of the circles are proportional to the coverage. The high-frequency smoothed values (described in detail in the Supplementary Methods) are plotted as solid lines. Dashed lines represent 95% pointwise confidence intervals. We used the region shown in Figure 3(c) to illustrate our statistical approach.